PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6:

A1 (11)

US

(11) International Publication Number:

WO 98/38326

C12N 15/86, 15/10

(43) International Publication Date:

3 September 1998 (03.09.98)

(21) International Application Number:

PCT/US98/03918

(22) International Filing Date:

28 February 1998 (28.02.98)

(30) Priority Data:

60/070,910

28 February 1997 (28.02.97)

Published

With international search report.

Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL,

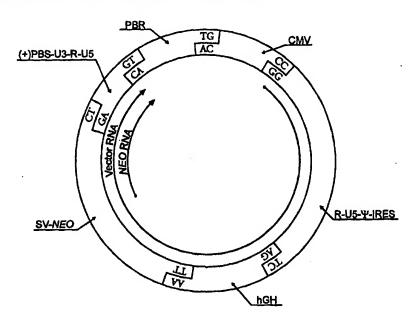
(81) Designated States: AU, CA, JP, US, European patent (AT, BE,

(71) Applicant (for all designated States except US): NATURE TECHNOLOGY CORPORATION [US/US]; 109 South 54th Street, Omaha, NE 68132 (US).

(72) Inventors; and

- (75) Inventors/Applicants (for US only): HODGSON, Clague, P. [US/US]; 109 South 54th Street, Omaha, NE 68132 (US). ZINK, Mary, Ann [US/US]; 109 South 54th Street, Omaha, NE 68132 (US). XU, Guoping [CN/US]; 109 South 54th Street, Omaha, NE 68132 (US).
- (74) Agent: MCCORMACK, Myra, H.; Mueting, Raasch & Gebhardt, P.A., P.O. Box 581415, Minneapolis, MN 55458-1415 (US).

(54) Title: SELF-ASSEMBLING GENES, VECTORS AND USES THEREOF



(57) Abstract

The invention relates to a method for directing the self-assembly of a gene or gene assembly having three and preferably six or more fragments in a directionally and spatially ordered fashion to produce a gene, gene vector or large nucleic acid molecule. The method can be used to create libraries, such as combinatorial libraries. In another embodiment of the invention a vector is described for the incorporation and screening of endogenous mouse promoter elements for the identification of cell-specific promoters.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav	TM	Turkmenistan
BF	Burkina Faso	GR	Greece		Republic of Macedonia	TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ.	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's	NZ	New Zealand		
CM	Cameroon		Republic of Korea	PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

SELF-ASSEMBLING GENES, VECTORS AND USES THEREOF

Field of the Invention

5 This invention relates to the construction and usage of synthetic genes for genetic engineering and gene therapy.

10

15

20

25

30

Background of the invention

This application claims the benefit of a provisional application U.S. Serial No. 60/070,910, filed on February 28, 1997, entitled "Self-Assembling Genes."

Recombination at the genetic level is important for generating diversity and adaptive change within genomes of virtually all organisms. Recombinant DNA technology is based upon simple 'cut-and-paste' methods for manipulating nucleic acid molecules *in vitro*. The pieces of genetic material, or DNA are first digested with a restriction endonuclease enzyme which recognizes specific sequences within the DNA. After preparation of two or more pieces of DNA, the ends of the DNA are further manipulated, if necessary, to make them compatible for ligation or joining together. DNA ligase, together with adenosine triphosphate (ATP) is added to the genes, ligating them back together. The genetic assembly containing an origin of DNA replication and a selectable gene is then inserted into a living cell, is grown up, and is positively selected to yield a pure culture capable of providing high yields of individual recombinant DNA molecules, or their products such as RNA or protein.

Significant improvements have been made to this technology over the last two and a half decades. Numerous enzymes, end-linkers and adapter molecules have been made commercially available, which facilitate in the construction of recombinant DNA molecules. By using two restriction enzymes with different single-stranded termini or blunt ends, it is possible to directionally assemble genes (forced cloning). This reduces the amount of screening required to determine orientation. Procedures have been automated for synthesis of single-stranded gene fragments up to 200 or more nucleotides in length by means of phosphoramidite chemistry, and the instrumentation is readily available through Applied Biosystems, Inc., Foster City, CA. Such single-stranded fragments can be joined by annealing overlapping complimentary phosphorylated strands, and by enzymatically filling in the ends with DNA polymerase and DNA precursors. In this way, multiple, overlapping, single-stranded fragments can be assembled into a larger, double-stranded superstructure.

Whole genes have been synthesized by similar methods. However, it becomes increasingly difficult to use synthetic DNA strands when making genes larger than approximately one kilobase. Using gene amplification methods (e.g. polymerase chain reaction (PCR), Mullis et al., U.S. Patent 4,683,195), together with synthetic oligonucleotides, it is possible to make biologically active, synthetic retro-vectors that are capable of RNA transcription, reverse-transcription, viral packaging, and integration into genomic DNA (see for example, Hodgson, WO94/20608). Hodgson, supra, also disclosed methods for cloning of transcriptional promoters into such a vector using traditional recombinant DNA technology.

5

10

15

20

25

30

Modified restriction enzyme sites, linkers, and adapters can change the primary or secondary structure of complex nucleic acid sequences thereby altering or obliterating a desired biological activity. For example, small mutations can drastically modify transcriptional promoters or change the reading frame of coding DNA. A logical goal of vectorology is to make exact constructs, without need of fortuitous restriction sites, adapters, or linkers.

Restriction endonucleases can be grouped based on similar characteristics. In general there are three major types or classes: I, II (including IIS) and III. Class I enzymes cuts at a somewhat random site from the enzyme recognition sites (see Old and Primrose, 1994. *Principles of Gene Manipulation*. Blackwell Sciences, Inc., Cambridge, MA, p.24). Most enzymes used in molecular biology are type II enzymes. These enzymes recognize a particular target sequence (i.e., restriction endonuclease recognition site) and break the polynucleotide chains within or near to the recognition site. The type II recognition sequences are continuous or interrupted. Class IIS enzymes (i.e., type IIS enzymes) have asymmetric recognition sequences. Cleavage occurs at a distance from the recognition site.

These enzymes have been reviewed by Szybalski et al. *Gene* 100:13-26, 1991. Class III restriction enzymes are rare and are not commonly used in molecular biology.

U.S. Patent No. 4,293,652 employed a linker with a class IIS enzyme recognition sequence to permit synthesized DNA to be inserted into a vector without disturbing a recognition sequence. Brousseau et al. (*Gene* 17:279-289, 1982) and Urdea et al. (*Proc. Natl. Acad. Sci. USA* 80:7461-7465, 1983) disclose the use of class IIS enzymes for the production of vectors to produce recombinant insulin and epidermal growth factor respectively. Mandecki et al. described a method for making synthetic genes by cloning small oligonucleotides using a vector (*Gene* 68:101-107, 1988). Expansion of a population of

oligonucleotides required synthesis, cloning excision and fragment purification. The oligonucleotides were used to create a complete plasmid.

5

10

15

20

25

Lebedenko et al. (*Nucl. Acids Res.* 19(24):6757-6771) illustrated the class IIS enzymes and PCR for precisely joining 3 nucleic acid molecules for convention sub-cloning using BamHI. Tomic *et al.* (*Nucleic Acids Res.*, 18:1656, 1990), reported a method for site-directed mutagenesis using the polymerase chain reaction and class IIS enzymes to join two nucleic acid molecules. Two overlapping PCR primers were used where the primers included class IIS recognition sites. The primers included a region of complementarity to the template DNA and include one to a few site-directed mutations. Stemmer et al. (U.S. Patent No. 5,514,568) employed overlapping primers with class IIS enzymes to amplify a plasmid and to introduce specific mutations into DNA leaving all other positions unaltered.

There remains a need for the ordering and assembly of complex genes to overcome the problems associated with sequential sub-cloning such as multiple purification steps, the potential for sample loss, and the like. Moreover there is a need for eliminating the use of prokaryotic hosts and for minimizing or avoiding the risks associated with bacterial contamination resulting from the use of bacteria as intermediaries in the cloning process. Further, there remains a need for efficient methods to assemble large nucleic acid molecules or many-fragmented nucleic acid assemblies with precision.

Brief Description of the Figures

Fig. 1A. provides one schematic of six double stranded DNA fragments, each terminus comprising a unique overhanging two-nucleotide sequence complementary to only one other terminus

- Fig. 1B. illustrates a three-piece ligation where 100% of the clones tested contained the predicted fragment order and desired fragment orientation.
- Fig. 2. illustrates the use of a class IIS restriction endonuclease (as one example, Bpm1), restriction endonuclease recognition site and the selection of cohesive overhanging ends.
- Fig. 3A. illustrates an exemplary retrotransposon-derived vector including a murine VL30 LTR (NLV-3) and packaging signal, an internal ribosome entry site (IRES) from encephalomyocarditis virus (EMCV), a gene encoding a green fluorescent protein (GFP), additional internal VL30 sequences (solid bar), SV40 early region promoter and Tn5

aminoglycosidase phosphotransferase (neo) gene, PBR322 plasmid origin of replication and a plus-strand primer binding site (VL30). An exemplary vector sequence is provided as VLBPGN (SEQ ID NO:1). Fig 3B is an illustration of an LTR with the insertion of a U3 (transcriptional promoter)region rescued by reverse transcriptase-polymerase chain reaction (RT-PCR). The promoter is amplified from the RNA of a cell expressing the VL30 U3 region. Complementary overhanging ends are created using class IIS restriction endonuclease digestion sites within the LTR and within the promoter. Fig. 3C provides the linear structure of a VL30 RNA transcript from a mouse cell with a U3 region near the 3'-terminus of the RNA molecule. PCR primers include a class IIS enzyme recognition site to amplify the U3 region from the RNA resulting in a double stranded DNA molecule. Cleavage with a class IIS enzyme (here *Bpm*I), results in a double-stranded DNA molecule with end complementary to a site in the vector of Fig. 3A.

5

10

15

20

25

30

Fig. 4A. is a schematic illustrating steps for assembling a combinatorial library of *cis*or *trans*-acting nucleic acid sequences for assembly and screening, useful for the rescue of
biologically active species. Fig. 4b is a diagram of a U3 (transcriptional enhancer and
promoter region of an LTR illustrating several sub-divisions of the transcriptional control
region, including a distal enhancer region, an enhancer repeat region, a medial promoter and a
proximal promoter. These regions have been described for other vectors in Hodgson et al.
(1996. "Construction, Transmission and Expression of Synthetic VL30 Vectors" in Hodgson
ed. *Retro-vectors for Human Gene Therapy*. RG Landes Company, Austin TX). Segments
of these regions are amplified using primers for highly conserved sequences. Highly
conserved sequences are determine based on a comparison of known VL30 sequences such as
provided in Fig. 4.2 of Hodgson, 1996, *infra*). The parts are joined by annealing and ligation
to provide an ordered assembly. Each construct is an allele or a representative of allelic
variation in the combinatorial library.

Fig. 5 discloses two transcriptional promoters that have been rescued from mouse VL30 RNA sequences isolated from a mouse T-helper cell library. These promoters were assembled into a vector andintroduced into retroviral helper cells and packaged into recombinant retrovirus for introduction into human T-cells. After transduction to human T cells, a β-galactosidase reporter gene was expressed from the T cell-derived promoters.

Fig. 6 discloses 10 biologically active mouse VL30 promoters obtained from mouse liver RNA. These promoters were introduced into the vector of SEQ ID NO:1. The vectors

were introduced into retroviral helper cells and then packaged into retrovirus where they were introduced into human liver cells. The cells expressed the green fluorescent protein.

Fig. 7 illustrates a similarity plot of nucleotide sequences found in VL30 U3 regions.

Fig. 8 illustrates a retro-vector comprising six double-stranded DNA fragments that were self-assembled into a circular structure using unique overlapping termini created using class IIS restriction endonucleases. Three templates and twelve primers were used in conjunction with three class IIS enzymes to make the six fragments that were ligated in a single step. The vector was efficiently self-assmebled and was effectively transmitted by both DNA transfection as well as by retroviral transduction of the self-assembled DNA, without molecular cloning through a prokaryotic host (see Example 2).

5

10

15

20

25

30

BRIEF SUMMARY OF THE INVENTION

The invention described herein provides seamless, directional, ordered construction of complex DNA molecules, vectors and libraries. More particularly, it enables gene constructs to be assembled with greater efficiency and precision, and it enables multiple gene fragments to be assembled in the correct order and orientation without disturbing the internal structure of the gene. The method utilizes *in vitro* assembly of nucleic acid fragments and relies upon the unusual ability of certain enzymes to digest nucleic acid molecules at pre-determined sites without disrupting the structure of the gene. It is especially useful for the construction of genetic vectors for gene therapy or genetic engineering of cells and organisms. A particular application of the invention is in combinatorial, or evolutionary genetics, where it enables a large number of non-random, self-assembled constructs to be screened simultaneously for function.

In a preferred embodiment of this invention, the invention relates to a method method for assembling a gene or gene vector comprising the steps of: a) designing at least 6 primers to produce to amplify at least three fragments in at least three separate polymerase chain reactions wherein each primer comprises at least one predetermined restriction endonuclease recognition site that recognizes a restriction endonuclease that cleaves at a distance from the recognition site, a sequence complementary to a template nucleic acid for amplification, and bases positioned at the restriction endonuclease cleavage site that are selected to be complementary to only one other overhanging created from enzymatic cleavage of the fragments; b) combining the primers with template nucleic acid and performing the

polymerase chain reaction to produce multiple copies of an amplified template fragment incorporating the restriction endonuclease recognition site; c) digesting the amplified template fragments with one or more restriction endonucleases that recognize the restriction endonuclease recognition site of the primers to create overhanging termini wherein each overhanging termini is complementary to only one other overhanging termini on another fragment; and d) combining the amplified and digested template fragments in a ligation reaction to produce a directionally ordered gene, nucleic acid fragment or gene vector.

5

10

15

20

25

30

In a preferred aspect of this embodiment, the restriction endonuclease is at least one class IIS restriction endonuclease and preferably, the class IIS restriction endonuclease is selected from the group consisting of: AlwI, Alw26I, BbsI, BbvI, BbvII, BpmI, BsmAI, BsmI, BsmBI, BspMI, BsrI, BsrDI, Eco57I, EarI, FokI, GsuI, HgaI, HphI, MboII, MnII, PleI, SapI, SfaNI, TaqII, Tth111II. Still more preferably, class II restriction endonuclease recognition sites (to be distinguished from class IIS restriction endonuclease recognition sites), linkers, or adapters are not used to create the gene or gene vector. In one embodiment, the product of the ligation reaction is introduced into prokaryotic or eukaryotic cells. Preferably, at least one template nucleic acid sequence is chosen from the group consisting of: transcriptional regulatory sequences; genetic vectors; introns and/or exons; viral encapsidation sequences; integration signals intended for introducing nucleic acid molecules into other nucleic acid molecules; retrotransposon(s); VL30 elements; or multiple allelic forms of a sequence.

In another preferred aspect of this embodiment, the method is used to generate combinatorial libraries of a target sequence. Preferably, the target sequence is part or all of a gene. In one embodiment, the gene encodes a protein. In one embodiment, the primers amplify allelic variants of part or all of a gene.

In still another preferred aspect of this embodiment, the product of the ligation reaction is passed between eukaryotic cells using a virus particle, by cell fusion, or by transfection. Preferably the product of the ligation reaction is not introduced into prokaryotic cells. Moreover, the method further comprises combining at least one screening or selection step to select the products of the ligation reaction. In one embodiment, the product of the ligation reaction is mutated during passage in cells in order to generate genetic diversity and preferably the product of the ligation reaction is mutated by homologous recombination during passage in cells.

In another aspect of this embodiment, the method is used to isolate and identify regulatory sequences from a cell. In another aspect of this embodiment, cells containing the product of the ligation reaction are selected for enhanced biological activity. Preferably, the cells containing the product of the ligation reaction are selected for tissue-specific, hormone-specific or developmental-specific gene expression. Also preferably, the ligation reaction is a circularized gene vector.

5

10

15

20

25

30

In another embodiment of this invention, the invention relates to a nucleic acid primer having a 5' and a 3' end to amplify a nucleic acid fragment for the ligation of at least two fragments comprising: a restriction endonuclease recognition site that recognizes a restriction endonuclease, wherein the restriction endonuclease cleaves at a distance from the recognition site and creates overhanging termini; a sequence complementary to a template sequence to be amplified to produce the nucleic acid fragment; at least two nucleic acid bases positioned at the restriction endonuclease cleavage site and that form an overhanging terminus after cleavage by the restriction endonuclease, wherein the at least two nucleic acid bases are selected to be complementary to only one other overhanging terminus on another fragment of the ligation; and an affinity handle on the 5' end of the primer. Preferably the primer further comprises an anchor to provide stability to the restriction enzyme at the restriction enzyme recognition site.

In yet another embodiment of this invention, the invention relates to a method for isolating and identifying promoters comprising the steps of: a) obtaining a vector comprising at least a portion of a promoter region from a retrovirus transposon LTR and having two non-complementary overhanging termini; b) designing at least two PCR primers to amplify at least one region of a retrovirus transposon LTR from template nucleic acid to produce at least one nucleic acid fragment wherein each primer comprises at least one predetermined restriction endonuclease recognition site that recognizes a restriction endonuclease that cleaves at a distance from the recognition site, a sequence complementary to a template sequence from a retrovirus transposon, and bases positioned at the restriction endonuclease cleavage site that are selected to be complementary to only one other overhanging terminus of the vector wherein the restriction endonuclease cleavage site is created from enzymatic cleavage of the fragments; b) combining the primers with template nucleic acid and performing a polymerase chain reaction to produce multiple copies of an amplified template fragment incorporating the restriction endonuclease recognition site; c)

digesting the amplified template fragments with one or more restriction endonuclease that recognize the restriction endonuclease recognition site of the primer to create overhanging termini; and combining the amplified and digested template fragment in a ligation reaction with the vector to produce a gene vector with an intact LTR sequence. In one embodiment of this aspect of the invention, the template nucleic acid is DNA or RNA. In another embodiment of this aspect of the invention, the method further comprises the step of sequencing the insert to identify the promoter sequence. In one embodiment promoter sequences of SEQ ID NOS:1-13 identified using the methods of claim.

5

10

15

20

25

30

Detailed Description of the Invention

In one embodiment of this invention, the invention relates to the seamless, oriented self-assembly of at least three DNA fragments having overlapping unique cohesive ends generated by the enzymatic cleavage of at least one restriction endonuclease that is capable of cleaving at a site distant to the restriction enzyme recognition site. Preferably the restriction endonucleases employed in this invention are class IIS restriction endonucleases. These enzymes recognize a predetermined group of nucleotides and cleave at a distance characteristic of the particular endonuclease from the recognition site. The term "unique cohesive ends" is used herein to refer to the notion that the cleavage site for the endonucleases of this invention can be manipulated to produce overhanging ends with unique termini selected by the investigator. The term "complementary" as used herein in reference to the overhanging ends of the fragments of this invention refers to standard complementarity recognized in the field of molecular biology. For example, the nucleotides sequence 5'-TAG-3' is said to be complementary to the nucleotide sequence 5'-CTA-3'. The term "PCR" is used generally to refer to the polymerase chain reaction and its variations, including RT-PCR as well as other gene amplification techniques employing primers.

In a first step for practicing one embodiment of this invention, a series of at least three overlapping fragments are created through the selection and creation of primers incorporating at least one class IIS restriction enzyme recognition sequence. The oligonucleotide primers of this invention are designed to amplify one or more nucleic acid fragments and comprise a sequence complementary to a target sequence for gene amplification, a recognition sequence for a restriction endonuclease that cleaves DNA at a distance from the recognition sequence (such as a class IIS restriction enzyme) and bases

5

10

15

20

25

30

positioned at the restriction endonuclease cleavage site that are preferably unique and complementary to only one other overhanging termini in the annealing/ligation reaction that generates the complex nucleic acid molecules. Optionally, the primers of this invention can include an "affinity handle for cleanup" at the 5'end. These sequences can be of any length, preferably at least about 6 bp and the sequences extend the primer in the 5' direction from the restriction enzyme recognition site. This extra length gives many enzymes greater stability and improved activity. In addition, the sequence can be used for recognition and removal of the ends of the primers (either undigested fragments or digested ends of primers) using complementary nucleotide sequences bound to a solid support (such as cellulose, nitrocellulose or silica). Incubation with, or passage over a column or support containing the complementary sequences can be used to remove the tags by allowing them to anneal or hybridize. The nucleic acid can then be eluted from the column. Adapters can also be used in this invention. For purposes of this invention, adapters refer to double stranded fragments containing an enzyme recognition site, according to this invention. The adapters are ligated to double stranded DNA molecules, creating a fragment analogous to a PCR fragment with similar sites derived from a primer. The primers or adapters can be prepared using a number of methods for synthesizing oligonucleotides known in the art. For example instruments for producing oligonucleotides are available from Applied Biosystems, Inc., Foster City, CA.

In one example, for the design of an oligonucleotide primer for use in this invention, the particular complementary bases that will form the site for hybridization of the primer to template (i.e., target DNA or RNA) are selected. A restriction endonuclease recognition site is selected followed by a number of nucleotides to be positioned between the recognition site and the cleavage site. The nucleotides of the cleavage site are selected to include overhanging regions formed from the restriction endonuclease cleavage that are complementary to the overhanging regions of an adjacent fragment in the annealing/ligation reaction.

The length of the primer used in this invention can vary, but preferably the primer length is up to about 80 bases and preferably up to about 50 bases. In addition the primers are preferably at least about 15 bases in length and preferably at least about 25 bases in length. The 5' region of the primer contains preferably at least about 6, preferably at least about 10 and still more preferably at least about 16-18 bases that are not complementary to the template DNA or RNA. Further, the primer incorporates a restriction endonuclease

recognition site preferably 5' to the region of complementarity and a restriction endonuclease digestion site preferably 5' to the region of complementarity or within the region of complementarity. There are a variety of restriction endonucleases that cleave at a distance from the restriction endonuclease recognition site of a DNA strand and a variety of enzymes that are commercially available from New England Biolabs are provided in Table 1.

Table 1. Restriction endonucleases useful in the construction of self-assembling genes

5

10

15

20

Enzyme:	Site size (bp):	Distance to overlap:	Size of overlap:	Overlap type:
Alw26 I	5	1-5bp	4bp	5'-Overhang
Bbsl	6	2-6bp	4bp	5'-overhang
<i>Bpm</i> l	6	16-14bp	2bp	3'-overhang
BsmBl	6	1-5bp	4bp	5'-overhang
BspMI	6	4-8bp	4bp	5'-overhang
<i>Bsr</i> Dl	6	0-2bp	2bp	3'-overhang
Eco571	6	16-14bp	2bp	3'-overhang
Fokl	5	9-13bp	4bp	5'-overhang
Hgal	5	5-10bp	5bp	5'-overhang
Hphl	5	8-7bp	1bp	3'-overhang
<i>MnI</i> I	5	7-6bp	1bp	3'-overhang
Plel	5	4-5bp	1bp	5'-overhang
Sapl	7	1-4bp	3bp	5'-overhang
SfaNI	5	5-9bp	4bp	5'-overhang

In addition to the enzymes provided in Table 1, other restriction endonucleases that cleave at a distance from their restriction endonuclease recognition site include, but are not limited to, AlwI, BbsI, BbvI, BbvII, BsmAI, BsmI, BsrI, EarI, GsuI, MboII, TaqII, Tth111II and their respective isoschizomers. These and other enzymes are known in the art and many are available from other manufacturers. The primers can be prepared to produce either 5'-overlapping ends or 3'-overlapping ends, as long as they are both are either 5'-overlapping ends or 3'-overlapping ends and are complementary to one other set of overlapping ends.

In the case of *Bpm*1 (see Example 1), the enzyme digests asymmetrically, 14-16 bp from the 3'-nucleotide of the recognition site. The resulting cleavage has a 3'-overhanging end of 2 bp. A second primer is then designed with a complementary

overhanging end, and it is used to generate the adjoining fragment terminus. At the opposite ends of the two fragments that are to be joined, similar complementary, overhanging ends are designed.

5

10

15

20

25

30

The oligonucleotides are then combined with template nucleic acid (either DNA or RNA, e.g., such as for reverse transcriptase polymerase chain reaction (RT-PCR)) containing bases complementary to at least a 3' portion of the primers (also referred to herein as "templates"). In one embodiment, the fragments are gene-amplified by PCR, RT-PCR or another gene amplification process using established PCR protocols such as those provided with PCR amplification kits, including those available from Perkin-Elmer Corp. (Emeryville, California). Preferably, the PCR products are analyzed by electrophoresis on a gel, such as an agarose gel and still more preferably the fragments of the predicted size are purified free of excess primers and small byproducts (such as by purification through a small column, such as a Oiagen™ column (Qiagen, Valencia, CA)). Following amplification or purification, the fragments are digested with the restriction endonuclease recognizing the restriction endonuclease recognition site in the primers. The digested fragments are then purified from the digested ends of the primers, preferably by preparative agarose gel electrophoresis. The fragments are combined, annealed and are ligated using standard hybridization and ligation conditions known for cloning (see Ausubel et al., Current Protocols in Molecular Biology, John Wiley & Sons, 1994).

Fig. 1A illustrates an example of a self-assembling gene construct (SEQ ID NO:1) comprising six fragments, each having unique overhanging dinucleotide ends. In this example, the ends of the fragments prepared by the methods of this invention are constructed using primers that include *BpmI* restriction endonuclease recognition sites. It will be understood by those of ordinary skill in the art that one or more other restriction endonucleases (such as those of Table 1) could similarly be used for the self-assembling product of Fig. 1A. In a preferred embodiment, the primers were created as described above and preferably the 3'ends of the primers are non-palindromic (i.e., non self-complementary) to prevent self-annealing of such fragments. Each fragment in this example preferably joins to only one other dinucleotide overhang in the annealing/ligation mixture, assuring ligation only to the intended fragment partner. An advantage of this strategy is that the formation of concatamers or multimers is minimal. The restriction endonuclease site is removed by

digestion with the restriction endonuclease, leaving the junction free of the extra DNA sequences associated with the site.

5

10

15

20

25

30

Using a single restriction endonuclease with a dinucleotide overhang (for example, using the enzyme BpmI) up to six pieces of genetic material can be joined together in a linear or circular form (such as a vector) without the need to perform sub-cloning procedures or detailed analysis of individual products because six unique combinations of dinucleotide overhangs create a directional clone with extremely high fidelity. With enzymes digesting single-base overlaps, only two fragments can be joined with positional and directional precision. With enzymes digesting three-base overlaps, 4³/2, or 32 fragments can be so joined in the correct order and orientation. Therefore, this invention also relates to the use of restriction endonuclease recognition sites that facilitate cleavage by restriction endonucleases with three-base overlaps and self-assembly gene constructs including 32 fragments. Alternatively, a combination of restriction endonuclease recognition sites for use with a combination of restriction enzymes that create two-base or three-base overlaps can be used. Each enzyme has its characteristic limits to self-assembly imposed by the size of the overlap. For example, there are sixteen dinucleotides, therefore BpmI fragments (which have two dinucleotide ends each) are limited to eight for the purpose of self-assembly; therefore in another embodiment of this invention an assembly comprising eight fragments is contemplated. However, four of the sixteen dinucleotides are palindromes. Use of these palindromic dinucleotides can create some infidelity in the annealing/ligation reaction. The enzyme HgaI has a five base overlap, and there are 1,024 pentanucleotide combinations. permitting 512 fragments to be ligated together directionally and in order (no palindromes). The fragments to be joined at a particular place are designed to have their cut sites aligned, so that the overlapping region fits together. In some cases, the target sequences will contain natural restriction endonuclease recognition sites for the enzyme that is being used, such as one or more internal BpmI sites. These sites have the potential to self-religate during vector or gene construction or they can be by passed by using a substitute enzyme in the primers (for example, Eco 571 can substitute for BpmI). Alternatively, these sites can be removed by sitedirected mutagenesis after consideration to the consequences of the mutagenized sequence to the gene or vector.

In addition to class IIS enzymes, class II restriction endonucleases can be used.

These enzymes have intrinsic methylation activity that affects the outcome in either a

negative or a positive way, depending on the purpose for which it is used. In a preferred embodiment, the methylation activity of class II enzymes is ablated by mutation or by genetic engineering to convert the enzyme to an effective class IIS enzyme to expand the repertoire of useful enzymes for this invention.

In another aspect of this invention, the primer design and target fragment sequence selection can be automated (see Example 5) using a computer to assist in the selection of unique overhanging ends that have complementarity only to the overhanging end of an adjacent fragment.

5

10

15

20

25

30

Therefore, this invention permits high-fidelity annealing and ligation of six or more fragments with unique overhanging termini complementary to a single other overhanging termini. Any multitude of combinations can be created by combining the type of overhanging termini that can be created. Moreover, if one is willing to sacrifice the fidelity of the reaction, a variety of combinations can be used to anneal a variety of fragment numbers. In these cases, some selection may be necessary, such as size selection of the resulting fragment based on electrophoretic migration or restriction endonuclease profiling, both methods well known to those of ordinary skill in the art

It is also necessary to have a high per-step efficiency (e.g., each step in the precess is performed with an efficiency of at least 80%) to effectively ligate large numbers of fragments without error. Where large numbers of fragments are used, the purity of the fragments becomes important. This means that for large numbers of fragments, the digested DNA fragments for annealing and ligation should be substantially pure. If undigested fragments, digested ends of primers, degraded or partially degraded molecules are present, they can decrease the purity and affect the fidelity of the product. Therefore, it is particularly desirable to ensure complete digestion of both ends of each fragment and to remove all of the digested ends from the fragments prior to including the fragments in an annealing and ligation reaction. The use of Qiagen columns for oligonucleotide removal prior to digestion is generally sufficient to permit efficient digestion of the fragments. Agarose gel isolation is desirable after digestion particularly where the product contains some fragments that do not appear to be full length. The use of an analytical gel before and after digestion helps in determining whether both oligonucleotide tags have been removed. The isolation of fragments from agarose gels preferably avoids the use of ultraviolet light and exposure of the

DNA to ethidium bromide is also preferably avoided. These methods can be avoided by running replicate lanes and staining only a portion of the gel.

5

10

15

20

25

30

The fragments and vector are then digested to yield fully complementary ends, and the fragments are preferably again purified, as described above (such as through a Qiagen column or by gel isolation). The purified fragments are ligated together in a test tube, under standard conditions, such as using bacteriophage T4 DNA ligase and ATP. Preferred ligations include at least 20µg/ml total DNA concentration in the ligation mix to favor intermolecular interactions, and an equimolar ratio of fragments to be ligated. Where a prokaryotic intermediary is used, the ligated assemblage is transformed into a bacterium, such as an *E. coli* host, and the colonies are: selected with a drug (such as an ampicillin, tetracycline, or kanamycin marker). The colonies can then be selected either by individually selecting colonies or growing a mass culture, such as where a vector library has been created. Restriction enzyme analysis can be used to determine the identity of individual constructs or to assess the validation of the combination of plasmids. The plasmids can then be grown up and used as needed.

In one embodiment of this invention, at least a portion of a vector is used as one of the fragments for the ligation of at least three fragments according to this invention. In one example, where a vector is used as one of the starting fragments, two restriction endonuclease recognition sites recognizing an enzyme that cleaves at a distance from the recognition site, such as at least one BpmI site, can also be introduced into the vector. This permits the vector to be digested with the restriction endonuclease to produce a product having ends complementary to two ends of the insert DNA fragments. The vector can be made by amplifying a plasmid or portion thereof using the primers of this invention. Thus, the vector can also be constructed to include a variety of restriction endonuclease recognition sites using a variety of restriction endonucleases, including a variety of class II restriction endonucleases. In some cases, the target fragments for amplification will contain natural restriction endonuclease recognition sites for the enzyme that is being used for the selfassembly, such as for example, a fragment that includes one or more internal Bpml sites. Care should be taken either to utilize the complementarity of the naturally occurring site to reform the fragment as it originally existed or to eliminate the restriction endonuclease recognition site using, for example, site-directed mutagenesis. Preferably, the restriction endonuclease recognition site is be substituted for a different enzyme (in the case of Bpml,

substituting *Eco*57I or *Bsr*DI) that has an equivalent structure at its ends. Two or more fragments of insert or two or more fragments of vector with at least one insert are amplified using primers according to this invention.

5

10

15

20

25

30

The exemplary enzyme, BpmI digests DNA 14-16 base pairs (bp) from the 3'nucleotide of the recognition sequence (RS). Thus, by placing the RS exactly 14-16 bp from the desired dinucleotide cut site, the practitioner tags the dinucleotide for ligation with another dinucleotide that is exactly complementary to it. Such a complementary dinucleotide can be inserted by using the same enzyme and RS to make another fragment which fits the first exactly, as illustrated in Fig. 1. Because there are sixteen possible dinucleotide combinations (including twelve combinations that do not have palindromic ends), it is possible to create up to six fragments with unique dinucleotides, and it is also possible to join them all together in a predetermined order and orientation (Fig 1A). In addition, the palindromic sequences (such as AT, CG, TA, and GC) could also be used, although inefficiency and incorrect ligation will result from the self-complimentarity of these sequences. It is furthermore possible and desirable to have three or more fragments joined in this way, such that the construct is circular as in Fig. 1, comprising a vector that may be grown in a bacterial and/or eukaryotic host cell. If the genetic construct is to be used as a vector, the vector should be designed to include a proper origin of replication to enable it to replicate in a particular cell. For example, a prokaryotic origin of replication such as a coliform plasmid origin of replication enables circular DNAs to be propagated in E. coli host cells. It is desirable to have at least one selectable marker, such as a neomycin marker that enables recovery of the clone through a selection process. It is also desirable, but not essential, to have two or more selectable genetic elements, to permit dual selection. For example, if one of the fragments contains a prokaryotic plasmid origin of replication, and another fragment contains a selectable marker, then the two fragments are both selectable, since the construct will grow in prokaryotic cells in the presence of a selection drug (such as ampicillin) only when both fragments are present. Drug selection can be combined with the methods of directed self-assembly to assure a high percentage of correct products. Because of the unique complementarity of the fragments, each contributes a selectable element that leads to recovery of a high percentage of correct products.

For prokaryotic vector construction, at least one fragment should contain a prokaryotic origin of replication and one fragment should contain a drug resistance marker

gene. However, an advantage of the methods of this invention is that the construct can be introduced directly into eukaryotic cells. Here no plasmid origin of replication is necessary and no prokaryotic selectable marker or other prokaryotic nucleic acid sequence is necessary. In cases where the vector is subject to regulatory approval or where optimal gene function is necessary, it may be undesirable to include prokaryotic sequences, such as extraneous plasmids or expressed prokaryotic fragments particularly if the sequences contain immunostimulatory sites that can lead to activation of the intracellular immune system and inactivation of a gene product (see Krieg et al., *J. Lab. Clin. Med.*, 128:128-133, 1996) or to avoid risks of endotoxin contamination. Moreover, the use of self-assembled product, according to the methods of this invention saves labor and time involved in the screening process.

5

10

15

20

25

30

Thus, in a preferred embodiment of the invention, the nucleic acid fragments are self-assembled in vitro, and are transferred directly into eukaryotic cells, by transfection, injection, or other methods known in the art. In one embodiment the cells receiving the assembled product of this invention are helper cells for recombinant virus assembly (including, but not limited to retroviral helper cells for retroviral or retrotransposon vectors, adenovirus helper cells for adenovirus vectors or herpes simplex virus helper cells for herpes simplex vectors). Alternatively, the assembled product can be introduced into cells along with a helper virus or the assembled product can be introduced into target cells for direct expression. The assembled product can be a vector, a minichromosome vector, a portion of a chromosome, or the like. In the preferred case of a retroviral vector, the genes are first transfected into a first helper cell line (such as ecotropic helper cells, GP+E86 (Markowitz et al. J. Virol. 862:1120-1124, 1988). The retrovirus-containing supernatant from these cells is then filtered (0.45mm Nalgene filters) preferably 48-72 hours after transfection and the filtrate is transferred to a second complementation retroviral helper cell line (such as PA317 retroviral helper cells, Miller et al., Mol. Cell. Biol. 6:2895-2902, 1986). After an additional 48 h, the second helper cell line is selected with the marker drug (such as the drug G418 for the selectable neomycin (neo) marker gene), until only drug-resistant cells remain. These cells contain stably integrated vectors that can be used to repeatedly transduce human cells. Advantageously, in the case of adenovirus vectors or other large eukaryotic -derived vectors including eukaryotic virus-derived vectors, it may be impossible to propagate them in prokaryotic hosts. The gene self-assembly method of the instant invention provides an

alternative to *in vitro* recombination method of gene construction by permitting large constructs to be constructed.

5

10

15

20

25

30

One advantage of introducing the assembled product of this invention into a helper cell line to produce recombinant virus for the introduction of a gene or nucleic acid complex into a cell is that the assembled product will be auto-selected by the cells during the packaging process. Therefore, even where the overhanging termini have palindromic sequences, where there is more than one (but preferably less than four) unique complementary matches for a particular overhanging termini, or where concatamers have formed, only the correct or functional assembled products are expressed, transmitted, and assembled into virus. When the virus is then introduced into cells, the use of a reporter gene or another selectable marker provides yet a second layer of security for the selection of cells containing a properly assembled construct. For example, where a retrovirus helper cell line is used to produce a recombinant retrovirus containing the product of this invention (for retrovirus, RNA transcribed from the DNA product of the invention becomes packaged into the virus particle), a retrovirus-derived vector is transcribed as RNA and transmitted by packaging the RNA in a retrovirus particle. In order to be properly transmitted as a virus, the construct must be: 1) transcribed as RNA in a vector producer cell; 2) packaged into viral particles; 3) reverse transcribed into double-stranded DNA (in the recipient cell); and 4) integrated into the host chromosome. Each of these steps requires specific cis-acting sequences that must be correctly positioned within the vector. Thus, passage via retrovirus (or by other virus) is a means of auto-selection for the essential sequences.

In one application of the methods of this invention, the methods are used to rescue expressed sequences from RNA, or genomic sequences from cell DNA without disrupting the promoter sequences. Cellular transcriptional promoters are typically difficult to identify and isolate because they are generally not included in the RNA molecule and often extend over a considerable distance in a chromosome. One application of this invention relates to a promoter rescue technique that permits the entire promoter, or a fragment of a promoter to be isolated and cloned directly in to an expression vector without disruption of the flanking sequences. Promoter rescue techniques are known and include WO 94/20608 to Hodgson.

In a preferred embodiment of the invention, transcriptional promoters are cloned in a transcriptionally active manner for the selection and identification of new and/or

of tissue or cell-specific promoters enabling them to be used, selected, or screened for activity directly. For example, Fig. 3 illustrates one example of the formation of a vector for the incorporation of promoter sequences and the ultimate identification of those sequences using an exemplary plasmid VLBPGN (SEQ ID NO:1) as provided in Example 3, with *Bpm*1 sites located within the locus of a retrotransposon (VL30) long terminal repeat (LTR). These methods preserve the structure and functionality of transcription factor response elements. The characteristic secondary structure of the LTR RNA remains very similar to the original LTR from which the promoter was rescued, thus preserving the important features of the original RNA/DNA molecule. Those of ordinary skill in the art will recognize that any of a variety of primers can be used with a variety of vectors and that the constructs of Figs 2 and 3 are exemplary and not limiting.

5

10

15

20

25

30

Fig. 2 illustrates the primers used to amplify the promoter insert (identified at a and c in Fig.2), and the insert region of the LTR (boxed), both of which can be digested at the same nucleotide position with Bpm1, to ensure a proper and seamless fit. In this example, after digestion of the vector, the two Bpm1 sites leave non-complementary ends (a 3'-CC overhang on one end, and a 3'-GC overhang on the other). Thus, the ends will not efficiently anneal or ligate to one another. However, the complementary termini of the insert serves as linkage, enabling the plasmid to be completed by ligation.

In the example illustrated in Fig. 2, the terminus on the 3'-side (GC) is palindromic. Palindromic termini are self-complementary and can therefore ligate to themselves or to an identical terminus facing the opposite way (forming concatamers in the opposite direction). Despite the presence of palindromic termini and despite the potential for reduced fidelity in the self-assembling process, a large percentage of clones obtained by inserting promoter sequences into VLBPGN were assembled correctly (20/23). These levels are reduced somewhat when three or more fragments are combined for self-assembly, according to this invention and preferably, the use of palindromic termini are avoided when even numbers of nucleotides are exposed as overhanging termini because with even numbers of nucleotides there is an axis of symmetry. As noted above, where five base overhangs are used there are 1024 possible combinations of five nucleotides [(4)⁵], yet none of them is palindromic.

The vector of Fig. 3 is an example of a particular type of vector that is known as a retrotransposon vector. Retrotransposon vectors are described and reviewed in Hodgson

et al., 1996 Retro-Vectors for Human Gene Therapy. RG Landes Company, Austin TX, chapter 5 and see US Patent 5,354,674 to Hodgson. This type of vector is derived from a mouse cellular retro-transposon element that has no essential viral or cellular genes, and that has little sequence similarity to a retrovirus. However, this RNA (known as VL30 [virus-like, 30S]) has all the necessary cis-acting structural elements (such as LTRs and primer binding sites) required for efficient transmission by a type C murine or primate retrovirus. Thus, it is a parasite transmitted by retroviruses that is also expressed as a cellular RNA in most mouse cells and tissues. This RNA becomes packaged into retroviral particles when the mouse cells become infected by retrovirus. The retrovirus then transmits the VL30 (or a VL30 vector) to the next infected cell (which can be a human cell). The RNA is then reverse transcribed and integrated into the DNA of the host cell.

5

10

15

20

25

30

Some advantages of VL30 vectors (over retrovirus-derived vectors) are: 1) lack of viral genes and other sequence homology that could lead to replication competent retrovirus (RCR); 2) ability to be expressed long-term *in vivo*; 3) a variety of LTR transcriptional promoters that can be expressed in various tissues and under the influence of various hormones and other stimuli; and 4) the ability to express genes in a number of cell types that are targets of gene therapy. An additional advantage is that VL30 parts can be switched with those of classical retrovirus-derived vectors. For example, the LTR or packaging signal of VL30 can be used in place of the equivalent retroviral signal. The ability to make mixed, or chimeric retro-vectors is a special application of gene self assembly technology.

Using a specific primer set, such as that shown in Fig. 2, or others, as taught in this invention, it is possible to amplify the U3 sequences expressed in the RNA of many different types of mouse cells. This is done using standard RNA isolation methods (Ausubel et al., supra), coupled with extensive digestion with ribonuclease-free dexoyribonuclease, to eliminate residual DNA. Thus, to obtain a promoter that is expressed in the liver, one isolates RNA from liver and uses an RT-PCR procedure, such as those known in the art, with the primers to amplify the desired promoters. Fig. 6 illustrates liver RNA-derived promoters obtained using the methods of this invention. However, the promoters can also be derived by conventional PCR from cDNA libraries (Fig. 5 illustrates T cell-derived promoters that were obtained in this manner). It is also possible to use the well-known hormonal and pharmacological inducibility of VL30 LTRs to find LTRs that are responsive to peptides,

hormones, and cytokines (for a table and description of VL30 pharmacologic responses (see Hodgson et al., 1996 Retro-Vectors for Human Gene Therapy. RG Landes Company, Austin TX, chapter 4, and Fig. 4.2). Examples of substances inducing various VL30 promoters to high levels include: epidermal growth factor, basic fibroblast growth factor, insulin, erythropoietin, glucocorticoid hormones, activators of cyclic 3'-5'AMP, and others. To rescue promoters with pharmacological responsiveness, cells or animals stimulated with the desired pharmacological agent are subjected to the RT-PCR procedure and the resulting U3 regions are cloned into a vector, (such as the exemplary VLBPGN) and are tested for inducibility. Standard RNA blotting procedures can be used before isolating VL30 promoters, to determine whether a particular drug or hormone causes induction of VL30 RNA expression in a particular mouse cell or tissue. After the promoter has been rescued, the vector is transmitted via retrovirus to the target cell (possibly a human equivalent of the mouse cell from which the promoter was rescued). After selection with the drug G418 (400-700 µg/ml, for 7-10 days) to select against cells not containing the vector, the target cell population is challenged with the pharmacological agent of choice. Reporter gene expression (in the example, GFP) or RNA expression, as determined by RNA blotting, can be used as an assay of gene inducibility by the agent (for exemplary gene expression methods, see Chakraborty et al., Biochem. Biophys Res. Commun. 209:677-683, 1995).

10

15

20

25

30

Using any specific primer set designed for use with VL30 retro-elements and using total cellular RNA from a particular mouse cell type as a template for RT-PCR, (using commercially available kits and methods therein) candidate promoter elements can be amplified. This method is useful for the identification of mouse-derived promoters and in particular the method is useful for the identification of cell-type specific or tissue-specific promoters from a mouse and for the selection of these promoters and the identification of tissue-specific or cell-specific promoters that function in human cells. Thus, these types of vectors and the methods for using these vectors permits the identification of promoters to permit controlled transcription of a foreign gene. The promoters, originally obtained from the mouse, can be used to effect tissue-specific or cell-specific expression in a human or animal liver cell such as a hepatocyte, or in a human blood cell such as a T-helper cell or in an erythrocyte (red blood cell). Methods are disclosed in Example 2 for the screening and selection of the promoters from a library of amplified promoter sequences. Other methods are well known to those of ordinary skill in the art. The specificity of the selected promoter

can be assessed, for example, by introducing a selectable marker under the control of the test promoter in question and introducing this construct into various cells to assess the ability of the promoter to selectively regulate expression.

5

10

15

20

25

30

The amplified fragments represent U3 promoter regions from any RNA species expressed in the originating cells and their abundance will be in approximate proportion to the number of expressed copies of RNA in the original mixture. Example 3 illustrates one example using a mouse T-helper cell cDNA library to produce amplified fragments representing U3 regions expressed in T cells. The vectors were efficiently expressed as RNA and protein in PA317 helper cells, and were transmitted by retrovirus into human T-helper cells, where they were integrated and expressed as protein in the form of a β -galactosidase reporter gene, as visualized by X-gal staining. The products of this experiment are provided in Fig. 5 and as SEQ ID NOS: 2 and 3 from T-helper RNA. The products of another experiment are shown in Fig. 6 as SEQ ID NOS: 4-13 from mouse liver RNA (by RT-PCR).

Examination of the different U3 sequences isolated from T cells and from liver revealed several things. First, the T cell U3 sequences were related to each other, as were the liver sequences. However, the two types of U3 sequences were quite different between the two sources (T-cell, Figure 5 and liver, Figure 6). Specifically, the liver sequences (Figure 6) appeared to be a closely related group, differing mostly by single point mutations, some of which may affect transcription factor binding sites. Some of the polymorphic sites included: a phorbol ester response element (VLTRE); a Rel/NFkb binding region, and a possible glucocorticoid response element (GRE). Some of these polymorphisms are illustrated in Fig. 6. The T cell-derived sequences (Fig. 5, SEQ ID NO:2 and 3), on the other hand, differed significantly in length, with SEQ ID NO:3 missing more than 120 bases (compared with SEQ ID NO:2) including putative binding sites for retinoids (RAR/RXR) and several elements contained within the enhancer repeat region (including a cAMP response element (VLCRE, or CREB/jun binding site), and putative serum response element (SRE, CARG, and NF1/IL6). SEQ ID NO:3 represented one out of five clones sequenced, while SEQ ID NO:2 represented four out of five. Possible sites of interactions between transcription factors and DNA can be observed by comparing the experimentally derived U3 sequences with those in Hodgson et al., (Retro-Vectors for Human Gene Therapy, 1996 Fig. 4.2 supra). In addition

to the deleted sequences of SEQ ID NO:2, there are a number of single base differences within the conserved regions of the two T cell-derived sequences.

5

10

15

20

Advantageously, a number of new VL30 promoter sequences (SEQ ID NOS: 2-13, *supra*) were identified using these methods despite the fact that VL30 RNA comprises only about 0.3% of cell mRNA represented in a cDNA library. Moreover, in each case, the cloned insert was isolated without the need to use linkers, adapters, or multiple cloning sequences such as those that are typically use for other library construction methods. The promoter sequences can be used in the vectors disclosed here to express inserted foreign genes or the promoter sequences can be substituted into other retroviral vectors, such as MoMLV-derived vectors or other VL30-derived vectors. Further, vectors containing the promoter sequences can be propagated in retroviral helper cells, such as PA317 (U.S. Patent 4,861,719 to Miller) or introduced into cells by chemical or physical transfection.

In another application of the methods of this invention, libraries of amplified sequences can be incorporated into vectors using two or more fragments and using the restriction endonucleases cleaving at a distance from their recognition sites. Preferably the vectors are created using six or more fragments and preferably greater than 10 or more fragments. For example, as applied to VL30 promoter sequences, because there are over a hundred VL30 retro-elements in the mouse genome, it is possible to amplify all of the promoter sequences *en masse*, and propagate them *en masse*, enabling screening by serial passage through helper cells (such as the PA317 helper cell line) or by means of a replication competent retrovirus, as illustrated in Examples 3 and 4. Conversely, the promoter region may be broken down into several sub-domains and permutations of each could be combined and screened to enhance the chances of generating a superior construct (Fig. 4B).

As an example of breaking a promoter region down into several sub-domains,

Fig. 7 illustrates a similarity plot of nucleotide sequences found in VL30 U3 regions. Plot similarity was performed using the Plot Similarity program (Wisconsin Sequence Analysis Package, release 8.1, Genetics Computer Group, Madison, WI). This program plots the running average of the similarity among the sequences in a multiple sequence alignment. The sequences compared were those found in Fig. 4.2 of Hodgson, 1996, chapter 4 (*infra*). That is, the plot discloses the degree of conservation of VL30 promoter sequences among known VL30 promoters. From the figure, it can be seen that conserved sequences (close to 100% conserved) can be used as primer binding sites to amplify the adjacent sequences by PCR.

An allelic mixture of three fragment sets is then created to make a combinatorial library of promoters that can be positively selected, such as by using retroviral amplification of the active sequences. This, used in combination with the Fig. 4.2 (Hodgson, 1996, chapter 4 supra) can be used to determine regions of high similarity. Regions of high similarity within the U3 region can be replaced with one another. Therefore, a library of permutations of these sections can be made by combining allelic pools obtained by amplifying the sequences from individual subsections, followed by ligating the subsections in the correct order using the methods of the instant invention for gene self-assembly. For example, sub-section 1 can include the distal enhancer (from the LTR 5'-end to the site of insert primer 2, see for example the region defined by the insert primers 1 and 2 (SEQ ID NOS 55 and 56 of Example 4). In this way, using a plot similarity (such as Fig. 7), within each sub-section, the primers position fragments within a region of nearly 100% identity. Degenerate primers can also be used in these experiments to account for multiple nucleic acid base combinations along a particular sequence. In each case, the primers preferably are designed to have a melting temperature that is compatible with the RT-PCR conditions being used, and the conditions should be those recommended by the manufacturer (preferably Perkin Elmer Corp., Emeryville, CA). In Example 4, a set of primers is given that can be used to amplify different U3 subsections, together with directions for assembling a combinatorial library.

10

15

20

25

30

It will be appreciated by persons of ordinary skill in the art that the methods of the instant invention can thus be used to make allelic libraries of a variety of genes. For example, different allelic portions of a gene can be combined in a predetermined order and orientation to produce combinatorial libraries, without the need for fortuitous restriction sites separating the parts in the original construct, and without perturbing the important sequences joining the parts using the methods of this invention.

In this invention primers are constructed as described above. However, for the generation of allelic libraries or more complex library constructs it may be helpful to include 5'tags into the 5' end of the primer. The purposes of the tag sequence are: 1) to provide extra nucleotides on both sides of the restriction endonuclease recognition sites (for more efficient digestion); and 2) to enable recovery of sequence tags or undigested fragments by means of an affinity reagent (such as silica, magnetic beads, or nitro-cellulose containing the complementary sequences) for purification. The use of an affinity reagent permits the digested ends to be purified away from the digested fragments. Furthermore, if any

undigested ends remain after thorough digestion, the affinity reagent will remove them, further aiding in the purification. In one embodiment, affinity purification of the digested fragments is used in place of gel isolation, eliminating possible damage caused by ultraviolet light as well as possible damage caused by dye (e.g., ethidium bromide) binding to the DNA.

5

10

15

20

25

30

It will also be appreciated that a number of other variations to the primer sequences can be employed. For example, as discussed above, the enzyme recognition site for an enzyme that digests outside of its recognition sequence is included in the primer, so that the DNA digest creates an overlapping end that is complementary to one other terminus to which it will be joined. The enzyme recognition site can be moved to any location within the primer so as to digest the DNA at the exact location desired. The primer can also be programmed with a novel enzyme recognition sequence to add any desired sequences between the two sequences to be joined or to incorporate a linker or adapter if desired. If the sequences to be amplified contain the enzyme recognition site of the primers, it may be necessary to switch to a different enzyme usage. The use of several different enzymes is possible and has been discussed above. As with other PCR procedures, after the initial primer selections have been made the primers are assessed for their ability to fold back on themselves or to create internal secondary structure. The primers are preferably modified to avoid palindromic sequences or the potential for self folding within a primer. Nucleic acid analytical software (such as the Wisconsin GCG package, Oxford Biomolecular, Oxford, UK) is available to perform this analysis and aid in the selection of alternative primers.

In addition, as with all PCR processes, it is necessary to determine the melting temperatures (T_m), and to adjust the annealing temperature of the PCR reactions to compensate for such temperatures. Finally, it is important to perform a sequence redundancy search, to determine whether the target sequence (the sequence complementary to the primer) is found more than once in the region to be amplified. If the sequence is repeated, it will be necessary to use a different primer in order to establish the single, correct priming site. Preferably, no more than 6-8 bases of incorrect target complementarity at the 3'-end of the complementary region is used and to allow a difference of at least 10° C between the T_m s of the correct and the incorrect target. The annealing temperature should always be at least 5°C lower than the T_m of the correct target and 5°C above the T_m of the incorrect target. Again, the necessary software and instructions are readily available from the cited sources (Wisconsin Gene Computer Group and Oxford Biomolecular, *supra*)

Next, a vector is constructed to include the appropriate elements for expression in the desired cell type. For example, the plasmid of Fig. 3A can be used for the creation of a promoter library or a vector can be created using a commercially available vector and primers to create a three or more fragment annealing and ligation reaction as provided above.

Preferably, the inclusion of a dominant negative selectable marker on the vector (e.g., the neomycin phosphotransferase gene, conferring G418 drug resistance) can be used to reduce the likelihood that cells without the vector are being maintained in culture.

5

10

15

20

25

30

Multiple allelic copies of DNA (cell derived or cDNA) can be amplified in separate reactions as a set of potential inserts with each set having its own unique overlap sequence following digestion with a restriction endonuclease, according to this invention. The fragments can then be ligated into an existing vector or in a single reaction of three or more fragments to form a combinatorial collection of potential alleles. For example, if six adjacent regions are amplified from five separate alleles, the number of combinations would be 56, or 15,625 potential combinations. The combinations can then be grown en masse, and selected in vitro or in vivo. A variety of screening strategies can be used in this invention and those of ordinary skill in the art will appreciate that the type of screen will match the type of library being generation. Therefore, for the promoter library, introducing members of the library into particular cell types to assess for expression in one or more cell types versus the absence of expression in another cell type is evidence of tissue-specific or cell-specific expression. For screening purposes, the libraries of this invention function like other libraries created through other methods. A variety of screening methods for a variety of libraries have been described in the art. For example, selective screens are reviewed by Hodgson et al. (1996, RG Landes Company, supra). Reporter protein production is well known in the art as is dominant selectable marker (e.g. drug) selection and selection by fluorescence activated cell sorting, antibody affinity selection, phage display selection (such as commercially available from Amersham, Milwaukee, WI), and the like can be used without detracting from this invention.

In this way, it is possible to isolate multiple forms of genes, gene fragments or regulatory regions such as transcriptional promoters or packaging signals (for example, in a retro-vector system). The individual constructs may then be tested *in vitro* or *in vivo* to further characterize a particular phenotype.

5

10

15

20

25

30

In one example the method is used to create a library of complementarity determining regions (e.g., allelic variations that give rise to antibody diversity) of antibodies or from receptors, including T-cell receptors, epitopes, antigens, ligands and the like. For example, where a library of T-cell receptors is created, the introduction of a vector designed to create a functioning T-cell receptor can be introduced into T cells or T-cell progenitors and the cells can be tested for their ability to bind to a particular test ligand. The ligandrecognizing cells can then be isolated from the ligand and grown in the presence of cytokines to produce specialized T cell clones. Where a library of antibodies or antibody fragments is created, the antigen reactive portions, for example, can be recombined in a vector containing the remaining portions of an antibody molecule to generate antibodies or antibody fragments in a cell. In other examples, the methods of this invention can be used to create allelic domains of receptor families (such as the steroid receptor super-family); libraries with related regions from peptide hormones; cytochromes P450; or other protein families that have shared domains or sub-sections with similar structures. The methods of the instant invention allow the joining of allelic sub-sections in an ordered fashion. In each case, it will be necessary to design primers, and to keep track of the uniqueness of joining overlaps and the presence of internal restriction sites as described above. While these will be different in each case, here are listed some general guidelines that are incorporated into the method of the instant invention.

As discussed above, although described as it relates to promoter libraries, libraries of other nucleic acid sequences can be created using the methods of this invention. These libraries include, introns and/or exons and/or functional domains libraries, libraries of potential alleles for a particular gene sequence, and the like. These sequences can be amplified from cell DNA or RNA using the primers of this invention and incorporated into a variety of vectors. For example, one vector of this invention, VLBPGN, has a portion of LTR removed and can be used to create a variety of libraries following digestion with *Bpm1*.

Selected or screened products of the combinatorial library can be used for gene expression, such as the promoters of Figs. 5 and 6. In addition, the exploitation of these sequences for the expression of a variety of genes, the LTR fragment containing the promoter can be joined to one or more functional retroviral packaging signals, internal ribosome entry sites, additional promoters, coding regions, processing sites, and the like.

Advantageously, there are almost no spatial constraints upon the joining of molecules by the method of the instant invention and other methods have not taken advantage of the combination of PCR to isolate genes or gene fragments; enzymes cleaving at a site distant from their restriction endonuclease recognition site to combine three or more fragments with precision; and, the use of unique overlapping non-palindromic termini to ensure fidelity of multi-fragment ligations. This combination permits the artisan to prepare complex gene constructions in one ligation step and does not require sequential sub-cloning into a vector or propagation in a prokaryotic host. Added to this the combination by these methods of fragment pools facilitates recombinatorial genetics.

5

10

15

20

25

30

The ability to recombine (in the correct order and direction) and screen a large number of allelic variants (whether as a simple library or as a combinatorial library), resulting in increased abundance (by amplification in the RNA, and subsequently in the DNA) is a special characterisitic of this invention. Particular advantages of this system are obtained when the methods of this invention are combined with retrovirus vector technology or other virus vector technology. For example, the combination provides a form of *in vitro* evolution whereby the passage of the library through virus and through cells selects functioning sequences and increases the abundance of the surviving RNA and DNA molecules.

For example, consider the consequences of screening several different promoters expressing RNA in a donor cell (*i.e.*, a cell producing virus particles), but at differing levels of RNA abundance. In the following example, the least abundant RNA species is expressed at 0.1 copy of RNA per cell, while six others are expressed at 1 copy, 10 copies, 100 copies 1,000 copies, or 10,000 copies, or 100,000 copies/cell, respectively. After a single passage, the DNA copy number in the recipient cells now reflects the approximate RNA copy number in the donor cells. These numbers are further amplified in the relative abundance of RNA species produced in the recipient cells. Disallowing for factors such as position effects, transcription factor depletion, etc., (which may be considerable), the same relative ratios of expression would be expected. Taking into consideration position effects, the disparity between abundance caused by changing insertion loci should average out. The most abundant RNA species after two passages is then many orders of magnitude more abundant than the least abundant.

Species:	RNA abundance: P=0	DNA copy no. P=1	RNA abun. P=1	DNA copy no. P=2	RNA abun. P=2
Α	0.1 copy/cell	0.1	0.01	0.01	0.001
В	1	1	1	1	1
С	10	10	100	100	1,000
D	100	100	10,000	10,000	10 ⁶
E	1,000	1,000	10 ⁶	10 ⁶	10°
F	10,000	10,000	10 ⁸	10 ⁸	10 ¹²
G	100,000	100,000	10 ¹⁰	10 ¹⁰	10 ¹⁵

Table 2. Enhancement of DNA and RNA copy number as a result of different RNA expression levels, after retroviral passage. P= (no. of passages). Numbers are interpreted as relative ratios within a column.

The present invention is able to efficiently create a library of RNA or DNA sequences whether or not they are in low abundance. The kinetics of screening for RNA abundance of a promoter can be appreciated best in the following discussion. For the purposes of this discussion, position effects have been ignored. An equation describing the kinetics of screening for RNA abundancy is:

(1)
$$R_{rel\chi} = A\chi / \sum A_{n-\infty}$$

5

10

15

20

25

The above equation (1) can be stated in plain English: The relative abundance of an RNA species χ ([R_{rel χ}]within a population of RNA molecules expressed in a single cell or within a population of cells) is equal to the RNA copy number of RNA species χ (A_{χ}) divided by the sum of the RNA copies of all RNA species present, including χ .

The relative abundance number of any given species changes as the number of passages change, according to the following approximation:

(2)
$$R_{\chi py} = D_{\chi p0} R^{p+1}$$

In the simplest of terms, equation two (2) can be expressed as: The abundance of RNA species χ after Y passages ($R_{\chi py}$) is equal to the initial abundance of the DNA for species χ at passage=0 ($D_{\chi p0}$), multiplied by the RNA abundance/DNA copy, raised to the power of the number of passages plus one. Thus, a typical RNA species that starts out as a

single copy of DNA, after zero passages (*i.e.*, in the donor cell) expresses 10 copies of RNA/cell. After one passage it is amplified at the DNA level to a relative ten copies (the same as the RNA abundance at P=0), and at the RNA level to 100 copies (10 copies per DNA copy). The reason for the amplification is that viral packaging and passage is based upon the number of RNA copies present in the donor cell. These calculations can be used to arrive at approximate abundance determinations for any given passage. The actual results of any given experiment, of course, will be biological rather than physical or mathematical. This means that other variables such as RNA efficiency of transmission and longevity, availability of transcription factors, experimental variation, *etc.* also come into play. The underlying purpose of the approximating equations, however, is to illustrate that RNA is amplified in DNA in proportion to the abundance of the template (RNA) within the cell.

5

10

15

20

25

30

The abundance of mRNA in cells can vary continuously from less than a copy per cell to nearly 100,000 copies/cell in actively transcribing, highly-specialized cells such as reticulocytes, the chicken oviduct, the silk moth silk gland, etc. Therefore, the spectrum of RNA abundance from 0-10⁵/cell is within the biological window of interest. For most practical purposes, such as biotechnological expression of genes in specific cells, only the higher end of this abundance range is desired. Therefore, using a viral selection system, as disclosed in this invention, it may be possible to disregard those species with less than a threshold level, such as <0.1 copies per cell. The selection through virus will lead to the recovery of the more abundant species. Furthermore, because the vector is likely to be the only considered sequence, it may be considered as a proportion of the whole of RNAs expressed in the target cell. The situation is more complex when a large number of permutations and combinations is generated, for example by self-assembling thousands or millions of fragments in a predetermined order using the self-assembly technique of the instant invention. Consider the assembly of allelic variants of four promoter subregions: distal enhancer, proximal enhancer, distal promoter and proximal promoter. If 100 varieties of each of the four groups were amplified and combined using the instant process along with a single vector, 108 resultant combinations could occur. However, a sufficient number of molecules to start out a combinatorial screening program might be a million. The problem can be simplified by considering these in groups as follows:

Table 3. Grouped abundance of RNA molecules derived from combinations.

5

10

15

No. of species in group:	RNA abundance:	Total No. RNA molec. at P=0:		RNA at P=2	RNA at P=3
9 X 10 ⁵ 2 X 10 ⁵ 2 X 10 ⁴ 1 X 10 ³ 1 X 10 ¹ 1	1 10 1,00 1000 10,000 100,000	2 X 10 ⁶ 2 X 10 ⁶ 1 X 10 ⁶ 1 X 10 ⁵ 1 X 10 ⁵	9 X 10 ⁵ 2 X 10 ⁷ 2 X 10 ⁸ 1 X 10 ⁹ 1 X 10 ¹⁰	9 X 10 ⁵ 2 X 10 ⁸ 2 X 10 ¹⁰ 2 X 10 ¹² 1 X 10 ¹³ 1 X 10 ¹⁵	9 X 10 ⁵ 2 X 10 ⁹ 2 X 10 ¹² 2 X 10 ¹⁵ 1 X 10 ¹⁷ 1 X 10 ²⁰
Sum Total:		6.6 X 10 ⁶	1.11 X 10 ¹⁰	1.01 X 10 ¹⁵	1 X 10 ²⁰

Thus, it follows that in the example population (Table 3) of over a million

constructs (equally represented in the DNA), a single construct expressing 105 copies of RNA per DNA copy will increase to approximately 99% of the total expressed RNA sequences in two passages. Using similar procedures in combination with drug and/or hormonal stimulation, and after consideration of the possible transcription factor binding sites within the sequence family (Figs. 5 & 6), it is within the intended scope of the invention to select for hormonal or pharmacological controls of transcription such as have been described herein. The factors contributing to the outcome are not only the input constructs, but recombinants and mutants as well. These secondary contributors to molecular diversity will be enhanced if multiple rounds of infections are allowed to occur, as oftentimes the difference between a particular transcription factor being able to bind (or not) may depend upon a single base change. Because viral infection is progressive and competitive, molecular evolution can be used to generate gene constructs de novo in the tissue culture dish in short time periods. Advantageously, the use primers to generate amplified fragments with uniquely complementary cohesive ends (i.e., that the ends will preferably only hybridize with the intended 5' and 3' fragments) to ligate three or more fragments as taught in this invention improves the potential for obtaining a diverse library.

Although the examples particularly point out a transcriptional promoter as the product of the process, the skilled artisan can appreciate that a particular selection technique can be applied to other *cis*- and *trans*-acting genetic sequences as well. Although a virus is used to propagate the selective advantage of a preferred embodiment, it can also be appreciated that any selective screen, such as drug selection, cell survival, phenotypic selection, cell sorting, antibody selection, and the like (see Ausuble et al., *supra*) could be

substituted without changing the intended scope of the invention. Likewise, transfection or cell fusion could be used in place of viral infection. Furthermore, substitution of different viruses, retrotransposons, or functional groups are likewise within the intended scope of the invention. The described embodiments are to be considered only as illustrative and not restrictive, and the scope of the invention is indicated by the claims rather than by the narrative description. All references and publications, cited herein, are incorporated by reference into this disclosure.

5

10

15

20

25

30

Like the embodiments detailed above, the method of library production is also conducive to assembly and transfer of genetic material directly into eukaryotic cells, saving the step of propagation in bacteria that is standard in bacteria. An advantage of direct transfer of the libraries of this invention to eukaryotic cells, including the exemplary retroviral vector producer cells, is that certain essential *cis*-acting structural features will be under positive selection (i.e., if they are not present, the molecule will be lost due to its non-functionality). As discussed above, it is often advantageous to eliminate bacterial and plasmid DNA sequences, endotoxin, and other bacterial contaminants by introducing the constructs directly into eukaryotic cells.

In addition to providing a method for constructing complex DNA molecules efficiently (as in the examples of three piece and six piece constructs), the methods of this invention permit the assembly of constructs that are larger than those conventionally propagated in E. coli. Examples of these types of vectors include adenovirus vectors, herpes simplex vectors and artificial minichromosomes. In order to insert genes into such vectors that are too large for conventional molecular cloning procedures, in the past it was often necessary to resort to in vivo recombination, wherein the genes of interest are cloned into a suitable vector and the flanking homologous regions are used to target the foreign genes to a homologous site within the larger viral or minichromosome vector. However, the methods of this invention permit PCR fragments of any size (up to the limits of PCR capability, 20-30 kb per fragment) to be joined together. Thus, it is feasible to precisely construct adenovirus vectors by amplifying larger sequences, and combining them by ligation. For example, several sections of adenovirus (5-10 kb each) can be ligated using the methods of this invention, up to for example, about 37 kb, and then transformed directly into human cells. Only the correctly recombined vectors are capable of replicating. Hence, the DNA is autoselecting. A similar procedure is used for generating herpes virus vectors, which are

approximately 150 kb. The precision of the methods of this invention permit non-essential-viral genes to be more easily eliminated from the construct. After transfection into appropriate cells, the DNA replicates and virus particles are formed.

5

10

15

20

25

30

Some special considerations apply to larger vectors, however. First, it is desirable to use enzymes that do not cut within the large DNA fragments. To prevent excessive fragmentation of the DNA by internal sites, it is desirable to use enzymes that cut rarely or infrequently, such as CpG-containing enzymes recognizing six bases, or enzymes such as Sap1, recognizing seven bases and digesting a three bp overhang (thus permitting up to 32 fragments to be joined in order). It is also desirable to avoid shearing the DNA once large segments have been joined by ligation. One method of avoiding shear is to add the transfection agent, such as Superfecttm reagent (dendrimers, Qiagen) or Lipofectaminetm (liposomes, Life Technologies, Gaithersburg, MD) directly to the ligation reaction, and then add the cells to be transfected to the mixture. This, or a similar method avoids the need to physically move the ligated DNA, and thus prevents shearing. Another method is to add a DNA condensing reagent (dendrimers, polycations [such as polyethyleneamine] histones or liposomes) directly to the DNA ligation reaction, and then move the DNA by pipette after it has condensed (thus reducing shearing of the DNA). Once inside the cell, viral DNA can replicate (as in the examples of partially replication-competent adenovirus and herpes simplex virus vectors).

Artificial minichromosomes have been under development for years. True artificial chromosomes require a centromere, at least one origin of DNA replication, and in the case of linear molecules, telomeric repeats at the chromosomal termini. In addition, to be very effective it is desirable to have a selectable marker gene, one or more therapeutic genes, and/or reporter genes.

In reality, the use of minichromosomes has been delayed by the inability to effectively manipulate the larger DNA molecules *in vitro*. Yeast and bacterial artificial chromosomes have been used with little success in mammals, and the addition of telomeres to the ends of linear chromosomes is also a special problem, as there is no prokaryotic host that can tolerate large linear DNA. The methods of this invention offers the opportunity to assemble human or mammalian minichromosomes *in vitro*, by using large segments (10-30 kb) of synthetic, gene-amplified DNA as ligation starting materials. For example, up to 32 Sap1 fragments (up to 30 kb each, containing the essential *cis*- and *trans*-acting sequences),

or 512 shorter *Hga*1 fragments can be combined using these methods. As with the other examples, several enzymes suitable for this invention (e.g., such as class IIS enzymes) can be combined (possibly with different termini lengths) to simplify the task. The methods of this invention also facilitate construction of telomeric repeats, because the constructs of this invention do not need to be circular. Thus, the methods of this invention can be used to make telomeres of any length, by adding additional segments onto the ends of molecules. One way to do this is using self assembling genes that employ a repeating overhang sequence (self-complementary molecule, such as AG-3' at one end, and CT-3' at the other end), permitting the telomeres to be lengthened to the extent desired by adding the required molar excess of the telomeric repeat-containing fragment. This technique gives the investigator some control over the relative length of the telomeres, although the self-complementarity indicates that many repeats will be lost due to self-ligation. This can be alleviated by using higher starting concentrations of DNA to favor inter-molecular ligations over intra-molecular ligations (e.g., >20 µg/ml starting concentration of DNA).

5

10

15

20

25

30

A two fold molar excess of telomeric fragments gives approximately twice the average length of telomere as a strictly 1:1 molar ratio of all fragments. By using a higher molar ratio of shorter telomeric repeats it is possible to give greater uniformity to the overall length of the molecules, which will vary from one terminus to the other. Thus, in addition to providing a way to build large molecules with precision, the methods of this invention provides for a way to control the telomere length (or potential life-span) of the artificial chromosome. To prevent damage during handling, the minichromosome DNA can be condensed with polycations, adenovirus particles, dendrimers, histones, or liposomes prior to transfection, similar to larger viral vectors.

The methods of this invention can be used to create recombinant virus. One example of this is an adenovirus vector self-assembling gene system. This system can include three parts: 1) vector: 2) helper virus; and 3) helper cells. The vector part is a self-assembling fragment set of at least three fragments comprising the essential cis-acting sequences (left and right inverted terminal repeats, which are the 103 bp at both ends of the genome that are required for replication [ITRs] and packaging sequences [Y, base pairs 194-358) and central 'baggage' area, comprising one or more self-assembling fragments including therapeutic genes, marker genes, and reporter genes. The baggage area is thus flanked by the cis-acting sequences in the vector. Because the synthetic oligonucleotide sequences

comprising the 5' and 3' termini of the helper virus are not phosphorylated, they will not ligate together creating multimers. Thus, the Ad5 vector region will assemble only into monomers. The helper virus part comprises all Ad5 trans-acting genes except for the ElA and ElB genes. The helper virus part has no cis-acting sequences, and it is amplified in several sections. In this preferred embodiment, the virus is amplified using primers that exclude the ITRs, packaging region and E1A&B genes. The helper virus is digested by Sap1 digestion, creating seven uniquely terminated fragments comprising the trans-acting viral genome, with dephosphorylated, blunt 5' and 3' ends on the terminating fragments. The primers are designed so as to amplify the internal virus sequences without changing them, except for the 5' and 3' ends of the virus. The PCR-amplified fragments are digested with Sap1 and are religated in their natural order after gel isolation and Qiagen column purification. The 5' end of the helper virus genome starts at 3.2 kb (in the E1A gene) so as not to overlap the vector sequences, which could otherwise cause replication competent adenovirus (RCA). Because the. 5' and 3' ends of the helper virus do not contain Sap1 sites, they remain intact after digestion with Sap1. Because the synthetic oligonucleotide sequences comprising the 5' and 3' termini of the helper virus are not phosphorylated, they will not ligate. Thus, the Ad5 helper virus genome assembles only into preferred monomers during ligation.

5

10

15

20

25

30

In a preferred embodiment, non-essential genes are deleted from the Ad5 genome by means of the method of self-assembling genes. In another preferred embodiment, the helper virus genome is approximately 30 kb after deletion of ElA, E1B and E3 gene sequences from the helper virus, and it is amplified as a single long fragment using the eLONGase Amplification System (Life Technologies or a similar strategy for creating long PCR fragments with high fidelity). It is not of great importance that occasional PCR errors may occur, because multiple copies of the Ad5 helper virus are transfected into target cells, thus providing trans-complementation. The helper cells are preferably 293 cells, a human kidney cell line expressing ElA and ElB genes (ATCC). The vector part and the helper virus part are combined in equimolar ratios after ligation has been performed separately on each fragment set. The Superfect protocol (Qiagen) is used to transfect the vector part and the he.lper part into the helper cells. The helper cells lyse, releasing high-titer adenovirus particles that are capable of infecting a variety of human cells. The resulting defective virus is incapable of forming RCA, and it transmits up to 34 kb of foreign genes in the baggage area. Unlike-conventional Ad5 vectors that require separate constructs for E. coli propagation of

insert genes, and recombination in vivo, the present vectors are relatively easy to make and provide a precise, safe alternative to first generation and second generation adenovirus vectors.

Exemplary methods for producing self-assembling vectors and genes are provided below. Further, the Examples provide methods for producing libraries of nucleic acid sequences using the methods of this invention. A number of nucleic acid sequences identified using the methods of this invention are described. The examples provided below are exemplary and not limiting. All references and publications provided herein are incorporated by reference into this disclosure.

10

15

20

25

30

5

Example 1 Three-Piece Gene Self-Assembly with 100% efficiency

Using 6 primers (SEQ ID NOS:24 and 63-67), three PCR fragments were amplified from templates VLMG (SEQ ID NO:22) and VLBPGN (SEQ ID NO: 1). PCR reactions were carried out using the hot start technique, according to the manufacturer's instructions (Perkin Elmer) using Pfu DNA polymerase (Stratagene). To amplify specific portions of the above templates, each primer contained a class IIS enzyme site capable of digesting a unique overhanging end that was complementary to only one other terminus in the subsequent ligation. The class IIS enzymes used were Bpm1 and Eco 57I (the latter was used to copy a fragment that contained an internal Bpm1 site). The reactions were carried out as follows: 1) the lower reaction was assembled according to the protocol for PCR Gems (Perkin Elmer); 2) the lower reaction was heated to 80°C, 5 min, then cooled to 4°C for 5 min; 3) the upper reaction was prepared according to PCR Gems protocol and was added to the lower reaction (separated by cooled wax). The primer concentration was 0.3 µM (final). The dNTP concentration was 200µM (final). 5 Units of Pfu polymerase was used. All fragments were amplified using the following conditions: 96°C, 45 sec; (then followed by 30 cycles of the following) 96°C 45 sec, 52°C 45 sec, 72°C, 6 min; then followed by a single incubation at 72°C for 10 min; then hold at 4°C. All fragments were successfully amplified. The PCR fragments were purified using the Qiaquick PCR purification protocol (Qiagen). The fragments were digested with an excess of the appropriate restriction enzyme (Bpm1 or Eco57I). The digested fragments were run on a 1% agarose gel and were excised using minimal irradiation from a hand-held 365 nm ultraviolet light. The fragments were purified

using the Qiagen Qiaquick Gel Purification Protocol. The fragments were ligated at an equimolar ratio at a concentration of >20µg/ml with T4 DNA ligase (Boehringer Mannheim) overnight at 4°C. Competent *E. coli* SCS110 cells (Stratagene) were transformed with the ligated DNA. Eight colonies were characterized by restriction enzyme analysis, and all eight contained the correct order and orientation of the three fragments. The experiment was repeated independently by another investigator, and the same result was obtained (8/8=100%). Thus, the procedure resulted in a high percentage of correctly assembled vectors.

This three-piece vector was VL Δ BP. The deletion extended from the distal enhancer region to the TATA box near the start of transcription. The deletion region was a pair of Bpm1 sites that permitted U3 sequences to be cloned into the insert.

One validated *E. coli* clone of VLΔBP was transfected into retroviral helper cells. After 48 h, the vector was transduced into amphotropic helper cells. After selection for two weeks with the drug G418, drug resistant colonies were grown up in a mass culture and the vector was transduced from the amphotropic helper cells into a human HT1080 cell line (ATCC, Rockville, MD). Surprisingly, even with a large deletion in the LTR promoter, the basal TATA box-containing VLΔBP was transmitted as a retrovector and was permanently inserted into the human cell line, thus establishing the validity of the self-assembly technique for the construction of functional eukaryotic vectors.

20

25

30

15

5

10

Example 2 Production of a Six Piece Self-Assembling Expression Vector

Due to the high efficiency of the gene self assembly process for the three piece assembly, a complex vector containing six fragments was constructed. The results here were extended to determine whether such a self-assembled vector would also have biological activity in human cells without being cloned and grown in a prokaryotic cell.

Six fragments were individually constructed by PCR using three different templates and twelve primers (as illustrated in Fig.8). The primers used three different class IIS enzymes. The enzymes were chosen so as to give 2 base pair, 3'-overhanging ends. Three enzymes were used in order to avoid the use of enzymes that had additional sites internal to the fragments being amplified. Thus, Bpm1 was used unless there was an internal Bpm1 site. If such a site existed, Eco57I was used. If there was also an internal Eco57I site, then BsrD1

was used. However, it is alternatively possible to use an enzyme such as *Eam*11041, where the *Eam*11041 sites in the primers are unmethylated (therefore susceptible to digestion by the enzyme), and wherein the ^{m5}dCTP analog of dCTP is used in the PCR reaction, methylating all internal sites (and protecting them from digestion by *Eam*11041), as suggested by Padgett and Sorge, 1996, *supra*, and incorporated herein by reference.

5

10

15

20

Using 12 primers, 6 fragments were amplified from 3 templates: pBK-CMV (SEQ ID NO:26), pVLMB (SEQ ID NO:23) and pVLOVhGH-900 (SEQ ID NO:21). Fragment 1 was amplified from pBK-CMV using primers 1 and 2 (SEQ ID NOS:31 and 32). Fragment 2 was amplified from pVLMB using primers 3 and 4 (SEQ ID NOS:33 and 34). Fragment 3 was amplified from pVLOVhGH-900 using primers 5 and 6 (SEQ ID NOS:35 and 36). Fragment 4 was amplified from pVLMB using primers 7 and 8 (SEQ ID NOS:37 and 38). Fragment 5 was amplified from pVLMB using primers 9 and 10 (SEQ ID NOS:39) and 40). Fragment 6 was amplified from pVLMB using primers 11 and 12 (SEQ ID NOS:41 and 42). PCR reactions were carried out using the hot start technique, according to the manufacturer's instructions (Perkin Elmer Ampliwax PCR GEMS 100). The lower reaction was heated to 80 ° C for 5 min, then cooled to 20 °C for 5 min. The upper reaction was prepared according to PCR gems protocol and was added to the lower reaction (separated by cooled wax). The primer concentration was 0.3 micromolar (final). The dNTP concentration was 200 µM (final). 5 U of Pfu polymerase (Stratagene) was used per reaction. 100 ng of template was used for each reaction 14 rounds of PCR amplification were used to reduce mutagenesis of the templates. The PCR cycling protocol was 96 °C 45 sec; then two cycles of (96°C 45 sec, 52°C 45 sec, 72°C 6 min); then 12 cycles of (96°C 45 sec, 58°C 45 sec, 72°C 6 min) followed by a 72° C soak for 10 min, then to 4°C hold.

The six PCR fragments were designed to self-assemble into a retro-vector after

digestion with the correct class IIS restriction enzyme (Fig. 8). After transfection into retroviral helper cells, the vector DNA is transcribed as RNA by means of the cytomegalovirus immediate early promoter (fragment 1). This promoter replaces the retroviral or VL30 LTR in this vector. The RNA transcript region begins with the R and U5 regions of the Moloney murine leukemia virus (MoMLV) LTR, the viral packaging signals

(Ψ) region of MoMLV, the packaging enhancer (Ψ+) region of mouse VL30 and the IRES region of EMCV fragment 2. Fragment 3 consists of the human growth hormone (hGH) cDNA-sequence. Fragment 4 consists of the SV40 virus early region promoter driving

expression of the neomycin phosphotransferase (neo) gene. Fragment five consists of the (+)-strand primer binding site of the MoMLV LTR, the U3 region of the MoMLV LTR, the repeat (or R) region, and a portion of the U5 region. Fragment 6 consists of the PBR322 plasmid origin of replication.

5

Fragment 1: CMV early region promoter

Template: pBK-CMV plasmid DNA (Stratagene, LaJolla, CA) Bpm1 (SEQ ID NO:26)

PCR primer 1 (SEQ ID NO:31)

10 GACTAACCTTGATTCCACTGGAGCCGTATTACCGCCATGCATTAGTTATTAATAG
PCR primer 2 (SEQ ID NO:32)

GACTAACCTTGATTCCACTGGAGTAATTGCGGCTAGCGGATCTGACG

Fragment 2: R-U5-Psi-Psi(+)-IRES Bpm1

15

25

Template: pVLMB plasmid DNA (SEQ ID NO:23)

PCR primer 3: SEQ ID NO:33

GACTAACCTTGATTCCACTGGAGACACTTGACCTCTACCGCGCCAGTCCTCCGAT
TGACTGAGTCG

PCR primer 4: SEQ ID NO:34

20 GACTAACCTTGATTCCACTGGAGGGATCCGCGCCCATGATTATTATCG

Fragment 3: human growth hormone (hGH) Bsr DI

Template: pVLCNOVhGH plasmid DNA (SEQ ID NO:21)

PCR primer 5: SEQ ID NO:35

GACTAACCTTGATTCCAGCAATGTCGGTTAGCTTGTTTCTTTACTGTTTGTC

PCR primer 6: SEQ ID NO:36

GACTAACCTTGATTCCAGCAATGTTAGGACAAGGCTGGTGGGCACTGG

Fragment 4: SV40 early promoter-neomycin phosphotransferase

Template: VLMB plasmid (SEQ ID NO:23)

30 PCR primer 7: SEQ ID NO:37

GACTAACCTTGATTCCACTGGAGGGTCGACCCTGTGGAATGTGTCAG

PCR primer 8: SEQ ID NO:38

GACTAACCTTGATTCCACTGGAGAATCTCGTGATGGCAGGTTGGGCGT

Fragment 5: MLV(+)PBS-U3-R-U5

Template: VLMB plasmid (SEQ ID NO:23)

PCR primer 9: SEQ ID NO:39

GACTAACCTTGATTCCACTGAAGAGATTTTATTTAGTCTCCAGAAAAAGGGGGG

PCR primer 10: SEQ ID NO:40

GACTAACCTTGATTCCACTGAAGCCCCCAAATGAAAGACCCCCGCTGACG

10

15

20

25

5

Fragment 6: PBR322 origin of replication

Template: VLMB plasmid (SEQ ID NO:23)

PCR primer 11: SEQ ID NO:41

GACTAACCTTGATTCCACTGGAGCCGGGACGGAATTCGTAATCTGCTGC

PCR primer 12: SEQ ID NO:42

GACTAACCTTGATTCCACTGGAGTTCTCGAGGCGGCGCATCTCGGCG

Procedure: The twelve primers were prepared by the following procedure: 1) oligonucleotides were synthesized with trityls off. After deprotection and lyophilization, the samples were resuspended in 5 microliters deionized formamide and loaded onto a polyacrylamide gel (12% polyacrylamide, 250V). The samples were excised under short wave UV irradiation and eluted overnight in 600 microliters of sample elution buffer (0.5 M ammonium acetate, 10 mM Mg acetate, 1 mM EDTA, 0.1% SDS). The contents were loaded onto a BioRad Chromatography column (Cat. # 732-6008) and centrifuged into an Eppendorf tube at low speed (2000 RPM, 5 min). After washing the column with 500 microliters TE buffer (10 mM Tris, 1 mM EDTA), pH 8.0 and recentrifugation (2000 RPM, 5 min), the pooled eluate was ethanol precipitated, washed with 100% ethanol, resuspended in TE buffer and quantitated by spectrophotometry of a small sample, which was then discarded.

Fragments were cleaned using the Qiaquick PCR cleanup procedure. The
fragments were digested with their respective class IIS restriction enzyme. The digested
fragments were run on 1% agarose gels, and the fragments were excised and cleaned using
the Qiaquick gel cleanup procedure. Fragments were combined in an equimolar mixture and

ligated overnight at 4° C with T4 ligase and ATP. An analytical gel was run with the ligated DNA, as well as with controls including unligated fragments and ligated fragments with a single fragment missing. As opposed to the controls, the complete ligation included bands equivalent to the full-length supercoiled monomer (referred to as GENSA 981, SEQ ID NO:29), as well as bands possibly representing multimers (up to six bands were observed).

5

10

15

20

25

30

In order to assess the efficiency of the method, eleven nanograms of DNA were transfected into SCS1 supercompetent cells. Thirteen kanamycin resistant colonies were harvested, and plasmid DNA preps indicated 10 out of thirteen that appeared to be the correct length. All ten gave the expected bands when digested with Pst1, SnaB1, and Bam HI. 1.35 ug of the ligated DNA was purified by phenol-chloroform-isoamyl alcohol extraction, followed by two extractions with chloroform-isoamyl alcohol, and was precipitated in ethanol. The DNA was washed in 70% ethanol and re-suspended in 50 µl of sterile phosphate buffered saline (for transfection). The DNA was transfected (using the Qiagen Superfect protocol) into HTam1 (amphotropic human helper cells). 24 h after transfection, the target cells were washed and fresh culture media was added. 48 h after transfection, the supernatant from the vector producer cells was filtered (0.45 µm, Nalgene) and transferred to PG13 helper cells (ATCC) and HT1080 human fibrosarcoma cells. This procedure was repeated after 72 h. 48 h after transduction, recipient cells were started on G418 drug selection (500 µg/ml). The appearance of G418 drug-resistant colonies on transduced PG13 and HT 1080 cells after 6 days of selection indicated successful transmission via retrovirus particles. The transfect HTam cells were also selected with G418. After six days of drug treatment, 45 colonies of resistant cells were counted. Thus, the six fragment gene assembly was effectively transmitted and expressed as either a DNA (transfection) vector or a retro-vector.

Example 3 Design and Construction of Single LTR Vectors

Background: In order to manipulate the interior of the VL30 LTR sequences using a promoter rescue technique, single LTR vectors were constructed. The mouse VL30 element NVL-3 was used as the starting material as it is constitutively and abundantly expressed in most mouse tissues. Single LTR vectors are circular and behave as if they contained two LTRs. Thus, in these vectors RNA transcription begins at the start of the R region (see Fig.

3B), and continues through the polyadenylation site after completing the second round of transcription of the R sequences (Fig. 3A). In previous studies, these vectors were expressed transiently in vector producer cells and the DNA did not integrate into cell DNA as a standard two LTR vector. Therefore, the vectors were usually passed to a second complementation helper cell line via retroviral transduction of the vector RNA transcribed in the first helper cell. This process resulted in the vector regenerating a correct (two LTR) structure upon integration into the recipient cell DNA.

5

10

15

20

25

30

Experimental method: The plasmid pNVL-3 (SEQ ID NO:25, kindly provided by Dr. J. Nortonm Manchester, UK), containing a complete copy of the NVL-3 (mouse VL30) genome (Adams et al, 1989), was digested with Xho1 (which cuts in the LTRs), releasing the 4.27 kb VL30 genome with one copy of the LTR. This fragment was circularized using T4 DNA ligase and ATP. The circular DNA was linearized by digestion with SnaBl, 187 bp from the 3'-LTR. A 2.3 kb fragment containing the SV40 virus early region promoter and the aminoglycoside phosphotransferase (neo) gene, together with the PBR322 plasmid origin of replication, was excised from the BAG retrovirus vector (Price et al., Proc. Natl. Acad. Sci. 84:156-160, 1987, kindly provided by C. Cepko, Cambridge, MA). BAG is also obtainable in a retrovirus helper cell line from American Type Culture Collection (ATCC), Rockville, MD by digestion with Xho1 and BamHI. This fragment was blunted with T4 DNA polymerase and dephosphorylated with calf intestinal alkaline phosphatase (CIP). The fragment was then ligated to the linearized SnaBI fragment of NVL-3. The resulting plasmid (containing a circularly permuted NVL-3 genome with the SV-neo-ori region) was designated VLSNO2 (SEQ ID NO:30).

In order to facilitate the switching of LTR sequences by means of the class IIS enzyme *Bpm*1, VLSNO2 was digested with *Bpm*1 (six sites). The region containing four *Bpm*1 sites was removed and replaced with a 19 bp linker (SEQ ID NOS: 1 and 52, see below), 921 bp beyond the LTR. The linker contained *Sna* BI, *Cla*1 and *Bam* HI cloning sites.

Linker (top strand): 5'-TACGTATCGATGGATCCGA-3' (SEQ ID NO:51)
Linker (bottom strand): 5'-GGATCCATCGATACGTAAG-3' (SEQ ID NO:52)

The remaining two of the *Bpm*1 sites had complementary ends, which permitted their ligation and resulted in eradication of all *Bpm*1 sites within the resulting vector VLSNO3 (SEQ ID NO:20).

In order to facilitate reporter/therapeutic gene function, a 3.7 kb fragment containing the internal ribosome entry site (IRES) from encephalocyocarditis virus, together with the β-galactosidase reporter gene, was excised from the plasmid pVLSAIBAG (kindly provided by Mr. James Grunkemeyer, Omaha, NE) by means of a partial digestion of the plasmid with *Bam* HI. This region was inserted into the *Bam* HI site of VLSNO3, resulting in the vector VLSNOSIB (SEQ ID NO:14).

A second reporter construct, pVLSNOG (5774 bp, SEQ ID NO:19) contained the green fluorescent protein (GFP, Clontech, Palo Alto, CA) gene was constructed by inserting a *Bgl2-Bcl1* fragment (800 bp) from plasmid pGFP-N1. This sequence, containing the GFP gene, was treated with mung bean exonuclease and inserted into the unique *Sna* B1 site of pVLSNO3.

In order to enhance GFP fluorescence from the reporter plasmid pVLSNOG, the serine-65 codon in the GFP gene was mutated into threonine by a site-directed mutagenesis procedure with the Transformertm Site-Directed Mutagenesis kit from Clontech. A *Bpm*1 site in the GFP gene (threonine-9) was mutated at the same time without changing the amino acid (ACT to ACA). The resulting plasmid was pVLSNOGM (SEQ ID NO:18).

An *Nco1-Xho1* fragment (585 bp) from plasmid pG1IL2EN (kindly provided by Dr. Steven Rosenberg, Bethesda, MD), containing the internal ribosome entry site (IRES) from encephalomyocarditis virus (EMCV) was inserted into the *Apa1* site upstream of the GFP gene in pVLSNOGM, resulting in pVLSNOGMI (SEQ ID NO:17). Both insert and plasmid fragments were blunted with mung bean exonuclease. One variant version of pVLSNOGMI with an IRES tandem dimer was also constructed and designated pVLSNOGMI2 (SEQ ID NO:16).

Oligonucleotides (SEQ ID NO:53 and 54) containing a splice acceptor (SA) of AKV virus (in bold) was inserted into pVLSNOGMI at the unique Sac 2 site just before the IRES, resulting in pVLSNOGMIS (SEQ ID NO:15).

Oligo: (SEQ ID NO:53)

5

10

15

20

25

30

5'-GGCCGCTAACTAATAGCCCATTCTCCAAGGTACGTAGC-3'

3'-CGCCGGCGATTGATTATCGGGTAAGAGGTTCCATGCAT-5'

(SEQ ID NO:54, bottom Oligo)

5

10

15

20

25

30

Recovery of LTR promoter sequences from mouse CD4+ T-helper cells

In order to facilitate the recovery of VL30 promoter sequences expressed in mouse T-helper cells, a mouse CD4+ T-helper cell cDNA library (Stratagene, San Diego, CA, Catalog # 937311) was screened by plaque hybridization. Approximately 2 x 104 bacteriophage λ-ZAP clones were plated on a lawn of E. coli cells according to the manufacturer's instructions. Two nylon filters were sequentially layered onto the lawn of E. coli cells and bacteriophage. The filters were hybridized to a 32P-labelled (Prime-It RmT Random Primer Labeling Kit, Stratagene), 4.2 kb internal Xho1 fragment of NVL-3 (containing the NVL-3 genome). 55 plaques (or approximately 0.3% of the total phage) reacted positively on both filters. 18 VL30 cDNA sequences were cloned from the plate, which was used to identify U3 promoters that are actively expressed in the RNA of mouse Tcells. Five of the 18 clones contained intact U3 sequences, representing four of one molecular species, named TH1 (SEQ ID NO: 2) and one of another species, named TH2 (SEO ID NO: 3) also provided in Fig. 5. TH1 contained approximately 120 bp more DNA than did TH2. Because TH1 was more abundant (4 out of 5 clones), the additional sequences in the enhancer region were implicated to be a possible reason for the stronger expression in mouse T cells. Examination of the known and putative transcription factor binding sites in the VL30 LTR (Hodgson, 1996, chapter 4, Fig. 4.2 supra) revealed several interesting features of TH1 and TH2. First, the extra sequences of TH1 that were missing in TH2 included an extra copy of the enhancer repeat region as well as a potential retinoid (RAR/RXR) binding site. Several transcription factor binding sites in the enhancer repeat region that differed between the two elements included: a cyclic 3'-5'AMP response element (VLCRE, a potential CREB/jun binding site), a serum response element (SRE), and a potential NF1/IL6 binding site (although there were additional sites for these factors in other enhancer repeats). These factors could possibly explain why VLTH1 appeared to be expressed at higher levels, both in the source cells and into transduced cells. Together, the VL30 sequences represented 0.3% of the mRNA expressed in the T cells, and TH1 appeared to be most abundant VL30.

Sequencing Primers:

(SK, SEQ ID NO:49) 5'-CGCTCTAGAACTAGTGGATC (20 mers, Tm 60°C). (T7, SEQ ID NO:50) 5'-GTAATACGACTCACTATAGGG (21 mers, Tm 60°C).

Seamless Rescue of T cell promoters using class IIS restriction enzymes 5

Two sets of primers containing offset Bpm1 restriction sites were designed and synthesized. One set was for amplification of the plasmid sequences, and another was for the amplification of the inserts.

Insert Primers: (Bpm1 site bold) 10

ITA (43 mer, Tm: 67.2 °C, SEQ ID NO:45)

CGATCCACTGGAGCTCGGAGCCCACCCCCTCCCATCTAGAGGT

ITB (43 mers, Tm: 66.3 °C, SEQ ID NO:46) 15

CGTCCTCCTGGAGAGCACAGGGTAGAGGAGTCTCGACGGTCAG

Vector primers: (Bpm1 site bold)

VLA (43 mers, Tm: 68.2 °C, SEQ ID NO:47)

CGCAACCCTGGAGACCTCTAGATGGGAGGGGGTGGGCTCCGAG

VLB (43 mers, Tm: 66.3 °C, SEQ ID NO:48) 20

25

30

GCAGGACCTGGAGCTGACCGTCGAGACTCCTCTACCCTGTGCT

To amplify vector sequences more efficiently, vector templates were shortened by deleting marker genes from vectors. pVLSNOSIB (SEQ ID NO:14) was cut with Kpn 1 and a 4201 bp fragment containing β-gal gene was removed. The remaining vector has 3923 bp.

The U3-promoter inserts (357 bp for TH1 and 240 bp for TH2) were PCRamplified from TH1 and TH2 promoters with primers ITA and ITB. The vector cassettes (~4.2 kb for pVLSNOSIB and ~3.7 kb for pVLSNOGMIS) were PCR-amplified from the shortened vector templates using primers VLA and VLB, (supra). The PCR-amplification was done with high-fidelity Pfu DNA polymerase from Stratagene (La Jolla, CA). The amplified products were gel-purified (1% agarose gel). The inserts were then cut with Bpm 1 to produce complementary ends. The vector cassette products were phosphorylated with PNK, then circularized with T4 ligase, and transformed into SCS 110 cells. Recovered plasmids were then digested with Bpm 1 and treated with CIP to produce complementary ends. Bpm 1 treated inserts and vector cassettes were ligated, and T-cell tissue-specific VL 30 vectors VLTH1 and VLTH2 were produced. The marker β -gal gene and GFP gene were put back into those vectors at the original unique sites Kpn 1 and Sal 1 respectively.

5

10

15

20

25

30

Transmission and expression of single LTR vectors and T cell U3 sequences

Vector DNA constructs were transfected into GP+E86 retroviral helper cells (Markowitz et al, 1988, *supra*) using the Lipofectamine protocol (Life Technologies, Gaithersburg, MD). The culture media from these cells (supernatant), containing defective transducing particles (72 h post-transfection), was transmitted to PA317 (Miller, US Patent, cited *supra*) amphotropic helper cells, using Lipofectamine to enhance transduction efficiency (Hodgson *et al.*, 1996. Synthetic Retrotransposon Vectors and Gene Targeting pp. 3-14, in: Felgner et al., eds. *Artificial Self-Assembling Systems for Gene Delivery*. American Chemical Soc. Books, Washington, D.C.). A similar procedure was used to transmit VLTH1 and VLTH2 to the PG13 helper cell line (Miller *et al.*, 1991. *J. Virol.* 65:2220-2224). 24 h post-transfection, the recipient cells were selected with the drug G418 (500μg/ml, 2 weeks) to enrich for stably transduced cell populations.

All of the single LTR vectors, including VLTH1 and VLTH2 were transmitted by this method, indicating that single LTR vectors can be used for promoter switching and yet revert to dual LTR vectors after a single passage. Vectors VLSNO2, VLSNO3, and VLSNOSIB were then titered on NIH 3T3 cells (using the PA317 vector producer cell lines). VLTH1 and VLTH2 vectors were titered on human HT1080 cells (PG13 cell lines). Surprisingly, all of the single LTR vectors were transmitted effectively. However the titers of stably transduced TH1 and TH2 cell lines were 5.5 x 10²-1.1 x 10³ TU/ml, compared to 0.4-3.0 x 10⁴ TU/ml for the VLSNO2, VLSNO3 and VLSNOSIB cell lines. Thus, switching from the NVL-3 transcriptional promoter (originally isolated from NIH 3T3 fibroblast cells) to VL30 promoters derived from T helper cells, appeared to have a negative effect on RNA expression in fibroblast cells, as determined by the transmissibility of the RNA.

In order to study the usefulness of rescued promoters as DNA transfection vectors (as opposed to retro-vectors), VLSNOSIB, VLTH1 and VLTH2 were also transfected into a number of cell lines (using Lipofectamine), including NIH 3T3, PA317, GP+E86,

PG13, HT1080, SW480 and HeLa (available from ATCC). RNA expression in these cell lines is shown in Table 4, wherein gene expression from the LTR promoter (as determined by β-gal staining) is normalized to VLSNOSIB (100).

Cell line:	NIH 3T3	PA317	GP+E86	PG13	HT1080	SW480	HeLa
Vector: VLSNOSIB VLTH1 VLTH2	100 39.3 28.6	100 18.7 7.1	100 0.1 5.5	100 21 11.5	100 25.5 46.8	100 156 82	100 156 156

Table 4. Transient expression of a β-gal marker gene by three VL30 promoters: NVL-3 (VLSNOSIB), VLTH1 and VLTH2. Cells were transfected using the Lipofectamine procedure. Total blue cells were counted from each well in 6-well plates, and the number of blue cells from VLSNOSIB was normalized to 100%.

10

15

20

25

30

5

The expression of both the VLTH1 and VLTH2 promoters was significantly reduced compared to VLSNOSIB in cell lines of fibroblastic origin, whereas in SW480 colorectal cancer cells and HeLa cells, it was comparable to or better than VLSNOSIB (the NVL-3 promoter). However, VLSNOSIB was expressed poorly in the non-fibroblastic cell lines, so a direct comparison was difficult to interpret. Unfortunately, the human T cell lines (Jurkat and MOLT4 [obtained from ATCC]) were not transfected by Lipofectamine, and they were poorly transduced by VLTH1 and VLTH2 retro-vectors. In the Jurkat and MOLT4 cells transduced with VLTH1 and VLTH2, only a small percentage (1-10%) of cells that were stably transduced by the vectors stained positively for β -gal expression. However, the marker gene (neo) continued to be expressed from an internal promoter, as evidenced by drug selection.

Taken together, the results demonstrated: 1) the ability of the promoter rescue technique to seamlessly capture functional transcriptional promoters from specialized cells; 2) the ability of single LTR vectors to introduce the rescued promoters into standard transducing vectors; 3) the ability of the rescued promoters to be expressed at differing levels in several different cell types, including T cells; and 4) screening and selection established the efficacy, or lack thereof, of individual promoter sequences.

Although the general method of promoter rescue was demonstrated by the foregoing experiments, the titers obtained from the sLTR VL30 vectors may not be useful where selection systems are not available.

Additional experimentation led to the development of a chimeric packaging signal, combining the essential packaging signal from Moloney murine leukemia virus (Ψ), and the enhanced packaging signal (Ψ +) from a mouse VL30 element. A vector embodiment of this packaging system is VLMB (SEQ ID NO:23). One advantage of the chimeric packaging system was the elimination of retroviral gag gene sequences that were present in previous high-titer MLV-based vectors (viral gag sequences contribute to the generation of replication competent retrovirus outbreaks). The titers of VLMB-based vectors ranged from approximately 1 x 10⁵ to 4 x 10⁶ TU/ml.

Construction of a cloning vector for promoter rescue

Using pVLSNOGMIS as a template, and primers (SEQ ID NOS:28 and 68), a 6.4 kb plasmid fragment was PCR amplified (Using Hot Start Ampliwax PCR Gems 100, Perkin Elmer). 30 cycles of PCR were performed by following the manufacturer's instructions, with the following input conditions: lower reaction, 80° C, 5 min., then add upper reaction and template, 96° C, 1 min. Each reaction vial contained 50 ng template, 0.5 μM each primer, 200 μM dNTPs and 5U (2μl) Pfu polymerase (Stratagene, LaJolla, CA). 30 repeating cycles of: 96° C, 45° sec; 50° C, 45 sec; 75 C, 1 min. A final incubation of 75° C, 10 min, then hold at 4° C. After amplification, the reactions were purified using Qiaquick PCR Purification Kits (Qiagen). The PCR products were digested with Pac1, heat inactivated (65° C, 20 min) and ligated together using T4 DNA ligase (overnight at 4° C in a 5 μl vol). The ligated DNA was transfected into SCS110 E. coli cells (Stratagene) with kanamycin (50 µg/ml) antibiotic added to the agar plates. The cells were dcm, dam (to prevent methylation of Bpm1 sites). The resulting plasmid, pVLBPGN (SEQ ID NO:1, Figs 2 &3) has a deletion in the U3 region of the LTR. A linker containing a central Pac1 site flanked by two outwardly-digesting Bpm1 sites occupies the site of the deleted U3 sequences. The Bpm1 sites enable the plasmid to be digested with Bpm1, resulting in two 2 bp 3'-overhanging ends that are complementary to the U3-derived RT-PCR inserts described below. The digested plasmid was purified free from the intervening linker sequences from an agarose gel after digestion with Bpm1, using the Qiaquick gel purification kit (Qiagen).

25

5

10

15

20

Procedure for amplification of liver U3 promoter region

Purified mouse liver total tissue RNA was purchased from Ambion, Inc., (Austin, TX). Total liver RNA was treated with RQ1 Rnase-free (Promega, Madison, WI). Using Perkin Elmer Gene Amp thermostable rTth reverse transcriptase RNA PCR kit (P/N N808-0069), the following conditions for RT-PCR were used: RT-PCR A 70° (hot start); RT-PCR B, 95°C, 60 sec, then 35 cycles (95°m 10 sec, 58°C, 15 sec) then a final 58°C incubation for 7 min, then 4°C and hold. Additional conditions were: primer concentration 0.15 micromolar, template 100 ng/reaction, dNTPs 200 micromolar (final) and MgCL₂ 3.5 mM(final). The primers for insert amplification were SEQ ID NOS:28 and 68)

The amplified U3 sequences were purified using Qiaquick. The pVLBPGN plasmid was digested with *Bpm*1, isolated from a 1% agarose gel and purified using the Qiaquick method. The purified U3 sequences were ligated at 1:2, 1:4 and 1:6 molar ratios of VLBPGN plasmid:insert using T4 DNA ligase and a 5 microliter reaction volume overnight at 4°C (100 ng plasmid: 16 ng insert = 1:1 molar ratio). 1 microliter of each ligation reaction was transformed into *E. coli* SCS 110 competent cells (Stratagene). 26 colonies were recovered in total. Out of 23 clones grown overnight in the presence of kanamycin, 20 had sequences that appeared to be mouse VL30 sequences, representing 10 different VL30 species (Fig. 6, SEQ ID NOS: 4-13). One of these (Hep 10, SEQ ID NO: 13) was transiently transfected into Hep G2 liver hepatocellular carcinoma cells. 48 h after transfection, intense GFP fluorescence was observed, indicating strong expression of the Hep 10 U3 promoter region.

Example 4 Creating a combinatorial library of mouse VL30 U3 sub-regions.

25

30

20

5

10

15

Using Fig. 7 and Hodgson, 1996, supra, Fig. 4.2 as a guide, the following three subregions of the VL30 U3 region were empirically established: Distal (1); medial (2); and proximal (3). Peaks of similarity were used to guide the following choice of primers: (+) primer binding site-5'-LTR boundary; ~80 bp (defines sub-region 1); ~80-210 bp (sub-region 2); ~210-430 (sub-region 3). The following primers were selected to amplify the vector VLBPGN or a similar VL30, NVL-3 LTR-containing vector:

P1 (going left from the 5'-end of the LTR to amplify the plasmid)

(SEO ID NO:55)

5

25

30

 ${\tt GACTAACCTTGATTCCACTGGAGTTTT(CT)(CT)ATTCTTCATTCCCCACTTC}\\ {\tt TTCTT}$

P2 (going right from the 3'-end of the promoter region to amplify the plasmid) (SEO ID NO:56)

GACTAACCTTGATTCCA**CTGGAG**AATCTGGACCAATTCTATATAAGCCTG TGAAAAATTT

The six primers selected to amplify the inserts are as follows:

Fragment 1, primer 1 (going right from the LTR terminus into U3) (SEQ ID NO:57)

GACTAACCTTGATTCCACTGGAGAAGAAGAAGTGGGGAATGAAGAA

Fragment 1, primer 2 (going left from the end of fragment 1) (SEQ ID NO:58)

GACTAACCTTGATTCCACTGGAGATCTCTAGATGGGAGGGG(GT)(CT)GGG

CTC

Fragment 2, primer 1 (going right from the left end of fragment 2) (SEQ ID NO:59)

GACTAACCTTGATTCCACTGGAGCTCGGAGCCCACCCCCTCCCATCT

Fragment 2, primer 2 (going left from the right end of fragment 2) (SEQ ID NO:60)

GACTAACCTTGATTCCACTGGAGGGAGGCCCTTATCTCAAAAATGTT

Fragment 3, primer 1 (going right from the left end of fragment 3) (SEQ ID NO:61)

GACTAACCTTGATTCCACTGGAGTCTAAGAACATTTTTGAGATAAGGGCC

T
Fragment 3, primer 2 (going left from the right end of fragment 3) (SEQ ID NO:62)
GACTAACCTTGATTCCACTGGAGTCACAGGCTTATATAG(TG)AAA

bears 100-200 copies of VL30 elements). Standard PCR procedures for *Pfu* polymerase are used. Fragments are amplified 35 rounds of PCR to obtain single-copy genomic DNA amplification. Samples of Qiagen column purified DNA are examined on analytical agarose gels to determine the approximate size. The remainder of each reaction is digested with the appropriate enzyme and run on an acrylamide or agarose gel. The digested fragments are purified by standard gel purification procedures and are ligated to the plasmid fragment at an equimolar ratio of the four PCR fragments (three inserts and one plasmid). The ligation mix

is transformed into *E. coli* SCS1 and is grown on kanamycin. The number of colonies is used to establish the size of the combinatorial library, and the pooled colonies are grown in *E. coli* and the DNA is harvested *en masse*. A dozen or more colonies are characterized by DNA sequencing to determine the approximate fidelity of the reaction. A library of 1,000 or more, but preferably 100,000 or more members is used for combinatorial screening procedures.

5

10

15

20

25

30

Screening the combinatorial libraries for expression in specific cell types using a replication defective helper virus

The U3 library DNA is transfected into the desired target cells in which expression is desired. Along with the library, approximately 25% of the total DNA should include retroviral helper sequences. The latter sequences can be a helper plasmid (such as pPAM3, Miller et al., US Patent 4,861,719). The virus is amphotropic, permitting it to infect most human cells. The RNA from individual clones that are transcribed in the target cells will be packaged into retroviral virions made by the helper virus, and the virions can be harvested as the cell free filtrate (0.45 mm) from the vector producer cells. This virus (containing the expressed sequences) can be transmitted to fresh target cells that do not contain helper virus. 48 hours after passage, the DNA form of the transcriptionally active clones will be integrated in the recipient cells, and these transcriptionally active loci will produce more RNA, and protein. After G418 drug selection to increase the proportion of cells expressing the vector sequences, helper virus DNA is again transfected into the recipient cells, transforming them into vector producer cells. The virus from these cells should contain increased amounts of the RNA from clones that are transcriptionally active in those cells. Passage of the virus is continued for two or three rounds to permit recombination and mutation to take place, enhancing the effect of in vitro evolution of promoters. The actual degree of enhancement attainable at each step is illustrated in Table 2 (supra). After several passages, the actual level of RNA expressed by several clones is determined by RNA blotting, or by the amount of a reporter gene expressed as protein (determined visually or by the appropriate assay). Because human cells do not naturally contain VL30 DNA or RNA, the sequences that remain in the human cells are those with the most transcriptionally active promoters. These sequences can be amplified and re-cloned using the methods of the instant invention, or they can be rescued by virus packaging, reverse transcribed by the endogenous reverse

transcriptase reaction, and grown as plasmids (due to their plasmid origin of replication and the selectable kanamycin marker gene).

In addition to using a replication defective helper virus, such as the clone pPAM3, it is also possible to use a replication competent retrovirus, such as Moloney murine leukemia virus to passage the library. For use in human cells, however, the virus should have a tropism that is compatible with human cells (gibbon ape leukemia virus and amphotropic [4070A] murine retroviruses are acceptable).

5

10

15

20

25

30

In addition to being useful for generating active transcriptional promoters *de novo*, a small variation on the above procedures may enable the isolation of hormone responsive promoters. In it, the cells are treated with the hormone (which could be a steroid, a peptide hormone known to affect the cells, a drug, a drug agonist or antagonist, etc.) during passage. After isolation of surviving VL30 vector-containing cells, individual clones of drug resistant cells are tested for reporter gene expression with and without drug treatment to determine relative protein expression. Likewise, RNA expression can be determined by blot analysis or a similar method. A useful list of known VL30 responses to pharmacological agents is listed in Fig. 4.2 of Hodgson, 1996, *supra*, and can be used as a guide to help assess the potential agents known to have an effect on VL30 transcription.

Once the transcriptional promoters with the known specificity have been obtained, they can be used to obtain expression of genes from a variety of types of vectors. For example, in addition to retrovirus particles, the promoters can be incorporated into all other major groups of vectors: adenoviruses, herpes simplex virus vectors, DNA transfection vectors, etc. It will be apparent to persons of ordinary skill in the art that similar combinatorial libraries can also be used to screen for other characteristics than transcription activity in a particular cell. For example, combinatorial libraries of complementarity determining regions (CDRs) of antibodies or T cell receptors can be so screened using antibody screening methods, such as the phage display screening method (Pharmacia, Milwaukee, WI). Thus, the methods of this invention, particularly the combinatorial simplicity of this invention is a significant improvement over many *in vivo* recombination methods including those of (Stemmer, US Patent 5,605,793; 1997) that have described for the production of CDR combinatorial libraries.

10

15

20

25

30

Example 5 Gene Assembly Line

From the above examples of 3 and 6 fragment gene self-assemblies, it is evident that assembly of genes by means of gene amplification, the use of offset restriction enzymes and incorporating unique, non-palindromic ends is a highly efficient process compared to conventional cloning methods. However, in addition to the considerations already discussed, it will be apparent to a person of ordinary skill in the art that the various procedures, protocols, methods and material of the instant invention become more difficult to use as the number of fragments increases. For example, if the efficiency of combining each fragment in an assemblage is 99%, then the overall efficiency of combining ten fragments will be 90%, the efficiency of combining 100 fragments will be 37%, etc. Therefore, a small drop in efficiency of any step or fragment, or a large increase in the complexity of the project, will be sufficient to reduce the overall efficiency. Fastidious procedures permit one to achieve success with more complex projects.

Foremost in its potential for inducing failure is human error in primer design where large numbers of fragments are used. Fortunately, the instant invention is suited to automation of most of the steps. This allows human input to be focused on design, analysis, and quality control. For the purposes of generating large vectors or chromosomes, it is desirable to provide an automated environment. One method to achieve this goal is a gene assembly line.

In a gene assembly line, multiple tasks are controlled by a machine or machines working together to increase speed and efficiency and to reduce human error. For example, computer aided design (CAD) and computer aided manufacturing (CAM) are incorporated and combined with the methods of this invention. The computers accept inputs in the form of template and primer sequences, together with preferences of regions to be copied and joined. The preferences include at least the sequences of the primer regions and information about the known restriction sites and maps of the sequences to be assembled, but ideally include the entire sequence. The preferences also include the number of sequences to be joined, the desired Tm for the primers, the list of potential restriction enzymes capable of offset digestion that are potential candidates for use in the assembly process, the desired end structures for each fragment terminus, a tag sequence (if any), whether circular or linear ends

10

15

20

25

30

are desired, and additional design considerations. The computer algorithm then searches the sequences to determine the candidate enzymes and specific primers that match the criteria of the input. Candidates for selection of unique non-palindromic overlaps are selected. The computer then posts selections or preferences for the type and order of end structures, the primer binding sites, their potential for primer-dimer and intra-molecular interaction artifacts, and the potential conflicts with repeat sequences within the templates that could lead to incorrect polymerization. Based upon the selections made by the operator, the computer then determine the T_m for each primer, and makes adjustments (with suitable inputs from the investigator) to achieve a suitable T_m for the appropriate DNA synthesis or gene amplification reaction. Ideally, the primers should have similar $T_{\rm m}s$ so that all amplification reactions can be performed at once with one set of amplification instructions. In reality, it may be difficult to do this with complex projects. The output of this portion of the program, which can be in a generic format, such as a Microsoft Excel spreadsheet is then downloaded to a computerized oligonucleotide synthesizer, such as the Applied Biosystems 3928 nucleic acid synthesizer. One advantage of using a computerized synthesizer is its robotic capability to de-protect and purify the oligonucleotides automatically. In addition this synthesizer can accept computerized input.

The quantity of individual oligos recovered is then determined spectrophotometrically. It is desirable to purify the oligonucleotides by high performance liquid chromatography or by polyacrylamide gel. In a preferred embodiment, the oligonucleotides and templates are then assembled robotically using an automated nucleic acid handling system such as the Qiagen BioRobot 9600. The BioRobot is capable of accepting input from a computer and can combine the gene amplification reactions based upon the assignments of templates, primer and reagents provided in the input. The assembled reactions are then amplified for example by PCR. In a preferred embodiment, the PCR heat block is incorporated into the robotic workspace and genes are assembled robotically but with minimal human intervention to change buffers, rearrange the platform, change programs, and the like. The resulting amplified products are also purified by the BioRobot or a similar robotic device. In a preferred embodiment, the robotic device uses Qiaquick cleanup procedures, or a similar method and then assembles restriction endonuclease reactions to digest the purified gene amplification products. The gene amplification products are loaded onto a gel and electrophoresed. Human intervention may be necessary to analyze the

products and excise the correct fragments from the gel. At this point, the results are assessed and missing or incorrect sized fragments are resynthesized. The robotic device is preferably used to purify the gel fragments using Quiagen or similar cleanup procedures. After spectrophotometric quantitation of the purified fragments, the robotic device is preferably used to assemble the ligation. Ideally the fragments are combined in an equimolar ratio of 1:1. However it is not necessary to use equimolar ratios in order to achieve gene self-assembly. For automated gene assembly, it may be desirable not to use equimolar ratios of input fragments, particularly if it simplified the task of quantitation. After ligation, the assemblies can be purified and ethanol precipitated or they can be added to the appropriate host cells. Automation aids in maintaining the sterility of the reaction.

5

10

15

20

25

30

Several additional considerations can assist in the construction of long genes using gene assembly. First the number of fragments and the length of constructs are limiting factors. In addition to maintaining high standards of purify of both the oligonucleotide primers and gene amplification products, it is important to keep the error rate low during copying. Thus, one can optimally start with 100 ng of template use only five rounds of gene amplification and finish with nearly 2 micrograms of product. This is more desirable for reducing errors than using a large number of amplification steps. It is also desirable to use a special copying enzyme such as Pfu DNA polymerase that has a low intrinsic error rate. Further it is desirable to use in vivo selection (in eukaryotic cells or tissues) rather than E. coli cloning to reduce the incorporation of errors into the vectors. For example, a viral vector such as an adenoviral vector or the retro-vectors of the preceding examples are auto-selecting. A single correctly-assembled adenovirus vector molecule, for example, leads to a lytic infection (the viral products of which are cloned by limiting dilution on the appropriate eukaryotic cells), even though it may be combined in a ligation mix with a large excess of incorrectly assembled molecules that are non-functional. Thus, it is not necessary to have a high efficiency, although high efficiency has been demonstrated in this system, in order to achieve success in making, for example gene therapy vectors.

For long fragments (3-30 kb), it is desirable to use enzymes and procedures that are designed or facilitate replication of long fragments, one such example is the eLONGase system (Life Technologies). This system can copy up to 30 kb on a fragment with proofreading. Considerations for long PCR are reviewed in Beck, 1998. (The Scientist 6 Janary, 1998, pp. 16-18).

Internal restriction sites are a potential problem, particularly with large constructs and can be overcome in a number of ways. Use of alternate enzymes, methylation of internal restrictions sites (such as by using methylated DNA precursors during synthesis to leave the sites in primers unaffected, incorporation of the internal sites into the construct (if they are non-palindromic), or mutagenesis of internal sites, are exemplary ways to deal with some of these issues.

5

10

15

20

For very large constructs, it is desirable to use enzymes such as SapI (recognizing 7 nucleotides and leaving a 3 bp overhang). This enzyme digests every 16,384 bp on average. There are 64 nucleotide triplet combinations, meaning that up to 32 fragments can be ligated in a circle using SapI. Fok1 and Hga1 are other examples of class IIS enzymes that are useful for making large constructs. Hga1 has 5 bp overhangs, permitting more than 500 Hga1 fragments to be ligated. Fok1 includes a Kozak ATG start codon. In a preferred embodiment, a Fok1 site is inserted at the PuXXATG start site of a cDNA encoding region. The cDNA is inserted in frame, providing a site for inserting and switching coding sequences within a vector.

It will be readily understood by those skilled in the art that the foregoing description has been for purposes of illustration only and that a variety of embodiments can be envisioned without departing from the scope of the invention. Therefore, it is intended that the invention not be limited except by the claims.

SEQUENCE LISTING

5	(1) GENERAL INFORMATION:
3	(i) APPLICANT: NATURE TECHNOLOGY CORPORATION, ET AL.
	(ii) TITLE OF INVENTION: SELF-ASSEMBLING GENES, VECTORS AND USES THEREOF
10	(iii) NUMBER OF SEQUENCES: 68
15	(iv) CORRESPONDENCE ADDRESS: (A) ADDRESSEE: MUETING, RAASCH & GEBHARDT, P.A. (B) STREET: 119 NORTH FOURTH STREET, SUITE 203 (C) CITY: MINNEAPOLIS (D) STATE: MINNESOTA (E) COUNTRY: USA
20	<pre>(F) ZIP: 55401 (v) COMPUTER READABLE FORM: (A) MEDIUM TYPE: Floppy disk (B) COMPUTER: IBM PC compatible (C) OPERATING SYSTEM: PC-DOS/MS-DOS (D) SOFTWARE: PatentIn Release #1.0, Version #1.30</pre>
25	(vi) CURRENT APPLICATION DATA: (A) APPLICATION NUMBER: Not Assigned (B) FILING DATE: 28-FEB-1998 (C) CLASSIFICATION:
30	(vii) PRIORITY APPLICATION DATA: (A) APPLICATION NUMBER: 60/070,910 (B) FILING DATE: 28-FEB-1997
35	(C) CLASSIFICATION:
40	(viii) ATTORNEY/AGENT INFORMATION: (A) NAME: MCCORMACK, MYRA M. (B) REGISTRATION NUMBER: 36,602 (C) REFERENCE/DOCKET NUMBER: 228.00010201
45	(ix) TELECOMMUNICATION INFORMATION: (A) TELEPHONE: 612-305-1225 (B) TELEFAX: 612-305-1228
	(2) INFORMATION FOR SEQ ID NO:1:
50	(i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 6225 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear
55	(ii) MOLECULE TYPE: DNA (genomic)
60	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:
	TGAAGAATAA AAAATTACTG GCCTCTTGTG AGAACATGAA CTTTCACCTC GGAGCCCACC 60
65	CCCTCCCATC TGGAAAACTC CAGTTATAAC TGGAGTTTTT CCTTTAAAAG CTTGTGAAAA 120 ATTTGAGTCG TCGTCGAGAC TCCTCTACCC TGTGCAAAGG TGTATGAGTT TCGACCCCAG 180

							_
	AGCTCTGTGT	GCTTTCTGTT	GCTGCTTTAT	TTCGACCCCA	GAGCTCTGGT	CTGTGTGCTT	240
	TCATGTCGCT	GCTTTATTAA	ATCTTACCTT	CTACATTTTA	TGTATGGTCT	CAGTGTCTTC	300
5	TTGGGTACGC	GGCTGTCCCG	GGACTTGAGT	GTCTGAGTGA	GGGTCTTCCC	TCGAGGGTCT	360
	TTCATTTGGT	ACATGGGCCG	GGAATTCGAG	AATCTTTCAT	TTGGTGCATT	GGCCGGGAAT	420
10	TCGAAAATCT	TTCATTTGGT	GCATTGGCCG	GGAAACAGCG	CGACCACCCA	GAGGTCCTAG	480
	ACCCACTTAG	AGGTAAGATT	CTTTGTTCTG	TTTTGGTCTG	ATGTCTGTGT	TCTGATGTCT	540
1.5	GTGTTCTGTT	TCTAAGTCTG	GTGCGATCGC	AGTTTCAGTT	TTGCGGACGC	TCAGTGAGAC	600
15	CGCGCTCCGA	GAGGGAGTGC	GGGGTGGATA	AGGATAGACG	TGTCCAGGTG	TCCACCGTCC	660
	GTTCGCCCTG	GGAGACGTCC	CAGGAGGAAC	AGGGGAGGAT	CAGGGACGCC	TGGTGGACCC	720
20	CTTTGAAGGC	CAAGAGACCA	TTTGGGGTTG	CGAGATCGTG	GGTTCGAGTC	CCACCTCGTG	780
	CCCAGTTGCG	AGATCGTGGG	TTCGAGTCCC	ACCTCGTGTT	TTGTTGCGAG	ATCGTGGGTT	840
25	CGAGTCCCAC	CTCGCGTCTG	GTCACGGGAT	CGTGGGTTCG	AGTCCCACCT	CGTGTTTTGT	900
25	TGCGAGATCG	TGGGTTCGAG	TCCCACCTCG	CGTCTGGTCA	CGGGATCGTG	GGTTCGAGTC	960
	CCACCTCGTG	CAGAGGGTCT	CAATTGGCCG	GCCTTAGAGA	GGCCATCTGA	TTCTTCTGGT	1020
30	TTCTCTTTTT	GTCTTAGTCT	CGTGTCCGCT	CTTGTTGTGA	CTACTGTTTT	TCTAAAAATG	1080
	GGACAATCTG	TGTCCACTCC	CCTTTCTCTG	ACTCTGGTTC	TGTCGCTTGG	TAATTTTGTT	1140
35	TGTTTACGTT	TGTTTTTGTG	AGTCGTCTAT	GTTGTCTGTT	ACTATCTTGT	TTTTGTTTGT	1200
33	GGTTTACGGT	TTCTGTGTGT	GTCTTGTGTG	TCTCTTTGTG	TTCAGACTTG	GACTGATGAC	1260
	TGACGACTGT	TTTTAAGTTA	TGCCTTCTAA	AATAAGCCTA	AAAATCCTGT	CAGATCCCTA	1320
40	TGCTGACCAC	TTCCTTTCAG	ATCAACAGCT	GCCCTTACTC	GAGCTCAAGC	TTCGAATTCT	1380
	GCAGTCGACG	GTACCGCGGC	CGCTAACTAA	TAGCCCATTC	TCCAAGGTAC	GTAGCGGGGA	1440
45	TCAATTCCGC	CCCCCCCTA	ACGTTACTGG	CCGAAGCCGC	TTGGAATAAG	GCCGGTGTGC	1500
43	GTTTGTCTAT	ATGTTATTT	CCACCATATT	GCCGTCTTT	GGCAATGTGA	GGGCCCGGAA	1560
	ACCTGGCCCT	GTCTTCTTGA	CGAGCATTCC	TAGGGGTCTT	TCCCCTCTCG	CCAAAGGAAT	1620
50	GCAAGGTCTG	TTGAATGTCG	TGAAGGAAGC	AGTTCCTCTG	GAAGCTTCTT	GAAGACAAAC	1680
	AACGTCTGTA	GCGACCCTTT	GCAGGCAGCG	GAACCCCCCA	CCTGGCGACA	GGTGCCTCTG	1740
55	CGGCCAAAAG	CCACGTGTAT	' AAGATACACC	TGCAAAGGCG	GCACAACCCC	AGTGCCACGT	1800
))	TGTGAGTTGG	ATAGTTGTGG	AAAGAGTCAA	ATGGCTCTCC	TCAAGCGTAT	TCAACAAGGG	1860
	GCTGAAGGAT	GCCCAGAAGG	TACCCCATTO	TATGGGATCT	GATCTGGGGC	CTCGGTGCAC	1920
60	ATGCTTTACA	TGTGTTTAGT	CGAGGTTAAA	AAAACGTCTA	GGCCCCCGA	ACCACGGGGA	1980
	CGTGGTTTTC	CTTTGAAAA	A CACGATACGO	GATCCACCG	TCGCCACCAT	GGGTAAAGGA	2040
65	GAAGAACTTT CACAAATTTT	TCACAGGAGT CTGTCAGTGG	TGTCCCAATT G AGAGGGTGA	CTTGTTGAAT GGTGATGCA	TAGATGGTGA A CATACGGAAA	TGTTAATGGG ACTTACCCTT	2100 2160

	AAATTTATTT	GCACTACTGG	AAAACTACCT	GTTCCATGGC	CAACACTTGT	CACTACTTTC	2220
	ACTTATGGTG	TTCAATGCTT	TTCAAGATAC	CCAGATCATA	TGAAACGGCA	TGACTTTTTC	2280
5	AAGAGTGCCA	TGCCCGAAGG	TTATGTACAG	GAAAGAACTA	TATTTTTCAA	AGATGACGGG	2340
	AACTACAAGA	CACGTGCTGA	AGTCAAGTTT	GAAGGTGATA	CCCTTGTTAA	TAGAATCGAG	2400
10	TTAAAAGGTA	TTGATTTTAA	AGAAGATGGA	AACATTCTTG	GACACAAATT	GGAATACAAC	2460
10	TATAACTCAC	ACAATGTATA	CATCATGGCA	GACAAACAAA	AGAATGGAAC	CAAAGTTAAC	2520
	TTCAAAATTA	GACACAACAT	TGAAGATGGA	AGCGTTCAAC	TAGCAGACCA	TTATCAACAA	2580
15	AATACTCCAA	TTGGCGATGG	CCCTGTCCTT	TTACCAGACA	ACCATTACCT	GTCCACACAA	2640
	TCTGCCCTTT	CGAAAGATCC	CAACGAAAAG	AGAGACCACA	TGGTCCTTCT	TGAGTTTGTA	2700
20	ACAGCTGCTG	GGATTACACA	TGGCATGGAT	GAACTATACA	AGTCCGGATC	TAGATAACTG	2760
20	TATCGATGGA	TCCGAAGGCG	GGGACAGCAG	TGCAGTGGTG	GACAGAAAGC	AAGTGATCTA	2820
	GGCCAGCAGC	CTCCCTAAAG	GGACTTCAGC	CCACAAAGCC	AAACTTGTGG	CTTTAATACA	2880
25	AGCTCTGTAA	ATGGTAAAAA	AAAAAAAGTC	TACACGGACA	GCAGGTATGC	TCTTGCCACT	2940
	GTACAGAGCA	ATATACAGAC	AAAGAGAACT	GTTGACATCT	GCAGAGAAAG	ACCTAAGATG	3000
30	CTGTGGCTAA	AAGAAATCAG	ATGGCAAATC	TAACCGCCCA	GGCATCCTAA	AGAGCAATGA	3060
30	TCCTGACAGT	CTGAAGACTA	TCAAGTTATA	GACAAATTAA	GACTGGTAAA	AAAAACCCTG	3120
	TATAAAATAG	TAAAAACTGA	AAAAAGAAAA	CTAGTCCTCT	CATGAGAAGA	CAGACCTGAC	3180
35	ATCTACTGAA	AAATAGACTT	TACTGGAAAA	AATATGTGTA	TGAATACCTT	CTAGTTTTTG	3240
	TGAACGTTCT	CAAGATGGAT	AAAAGCTTTT	CCTTGTAAAA	CGAGACTGAT	CAGATAGTCA	3300
40	TCAAGAAGAT	TGTTAAAGAA	AATTTTCCAA	GGTTCGGAGT	GCCAAAAGCA	ATAGTGTCAG	3360
70	ATAATGGTCC	TGCCTTTGTT	GCCCAGGTAA	GTCAGGGTGT	GGCCAAGTAT	TTAGAGGTCA	3420
	AATGAAAATT	CCATTGTGTG	TACAGACCTC	AGAGCTCAGG	AAAGATAAAA	AAGAATAAAT	3480
45	ААААСТСТАА	ACAGACCTTG	ACAAAATTAA	TCCTAGAGAC	TGGCACAGAC	TTACTTGGTA	3540
	CTCCTTCCCC	TTGCCCTATT	TAGAACTGAG	AATACTCCCT	CTTGATTCGG	TTTTACTCTT	3600
50	TTTAAGATCC	TTTATGGGGC	TCCTATGCCA	TCACTGTCTT	AAATGATGTG	TTTAAACCTA	3660
50	TGTTGTTATA	ATAATGATCT	ATATGTTAAG	TTAAAAGGCT	TGCAGGTGGT	GCAGAAAGAA	3720
	GTCTGGTCAC	AACTGGCTAC	AGTGAACAAG	CTGGGTACCC	CAAGGACATC	TTACCAGTTC	3780
55	CAGCCAGAGA	TCTGATCTAC	GATCCCCGGG	TCGACCCGGG	TCGACCCTGT	GGAATGTGTG	3840
	TCAGTTAGGG	TGTGGAAAGT	CCCCAGGCTC	CCCAGCAGGC	AGAAGTATGC	AAAGCATGCA	3900
60	TCTCAATTAG	TCAGCAACCA	GGTGTGGAAA	GTCCCCAGGC	TCCCCAGCAG	GCAGAAGTAT	3960
O.	GCAAAGCATG	CATCTCAATT	AGTCAGCAAC	CATAGTCCCG	CCCCTAACTC	CGCCCATCCC	4020
65	GCCCCTAACT TTATGCAGAG	CCGCCCAGTT GCCGAGGCCG	CCGCCCATTC CCTCGGCCTC	TCCGCCCCAT TGAGCTATTC	GGCTGACTAA CAGAAGTAGT	TTTTTTTAT GAGGAGGCTT	4080 4140
0.5	TTTTGGAGGC	CTAGGCTTTT	GCAAAAAGCT	TCACGCTGCC	GCAAGCACTC	AGGGCGCAAG	4200

							_
	GGCTGCTAAA	GGAAGCGGAA	CACGTAGAAA	GCCAGTCCGC	AGAAACGGTG	CTGACCCCGG	4260
	ATGAATGTCA	GCTACTGGGC	TATCTGGACA	AGGGAAAACG	CAAGCGCAAA	GAGAAAGCAG	4320
5	GTAGCTTGCA	GTGGGCTTAC	ATGGCGATAG	CTAGACTGGG	CGGTTTTATG	GACAGCAAGC	4380
	GAACCGGAAT	TGCCAGCTGG	GGCGCCCTCT	GGTAAGGTTG	GGAAGCCCTG	CAAAGTAAAC	4440
10	TGGATGGCTT	TCTTGCCGCC	AAGGATCTGA	TGGCGCAGGG	GATCAAGATC	TGATCAAGAG	4500
	ACAGGATGAG	GATCGTTTCG	CATGATTGAA	CAAGATGGAT	TGCACGCAGG	TTCTCCGGCC	4560
1.5	GCTTGGGTGG	AGAGGCTATT	CGGCTATGAC	TGGGCACAAC	AGACAATCGG	CTGCTCTGAT	4620
15	GCCGCCGTGT	TCCGGCTGTC	AGCGCAGGGG	CGCCCGGTTC	TTTTTGTCAA	GACCGACCTG	4680
	TCCGGTGCCC	TGAATGAACT	GCAGGACGAG	GCAGCGCGGC	TATCGTGGCT	GGCCACGACG	4740
20	GGCGTTCCTT	GCGCAGCTGT	GCTCGACGTT	GTCACTGAAG	CGGGAAGGGA	CTGGCTGCTA	4800
	TTGGGCGAAG	TGCCGGGGCA	GGATCTCCTG	TCATCTCACC	TTGCTCCTGC	CGAGAAAGTA	4860
25	TCCATCATGG	CTGATGCAAT	GCGGCGGCTG	CATACGCTTG	ATCCGGCTAC	CTGCCCATTC	4920
23	GACCACCAAG	CGAAACATCG	CATCGAGCGA	GCACGTACTC	GGATGGAAGC	CGGTCTTGTC	4980
	GATCAGGATG	ATCTGGACGA	AGAGCATCAG	GGGCTCGCGC	CAGCCGAACT	GTTCGCCAGG	5040
30	CTCAAGGCGC	GCATGCCCGA	CGGCGAGGAT	CTCGTCGTGA	CCCATGGCGA	TGCCTGCTTG	5100
	CCGAATATCA	TGGTGGAAAA	TGGCCGCTTT	TCTGGATTCA	TCGACTGTGG	CCGGCTGGGT	5160
35	GTGGCGGACC	GCTATCAGGA	CATAGCGTTG	GCTACCCGTG	ATATTGCTGA	AGAGCTTGGC	5220
33	GGCGAATGGG	CTGACCGCTT	CCTCGTGCTT	TACGGTATCG	CCGCTCCCGA	TTCGCAGCGC	5280
	ATCGCCTTCT	ATCGCCTTCT	TGACGAGTTC	TTCTGAGCGG	GACTCTGGGG	TTCGAAATGA	5340
40	CCGACCAAGC	GACGCCCAAC	CTGCCATCAC	GAGATTTCGA	TTCCACCGCC	GCCTTCTATG	5400
	AAAGGTTGGG	CTTCGGAATC	GTTTTCCGGG	ACGGAATTCG	TAATCTGCTG	CTTGCAAACA	5460
45	AAAAAACCAC	CGCTACCAGC	GGTGGTTTGT	TTGCCGGATC	AAGAGCTACC	AACTCTTTTT	5520
43	CCGAAGGTAA	CTGGCTTCAG	CAGAGCGCAG	ATACCAAATA	CTGTCCTTCT	AGTGTAGCCG	5580
	TAGTTAGGCC	ACCACTTCAA	GAACTCTGTA	GCACCGCCTA	CATACCTCGC	TCTGCTAATC	5640
50	CTGTTACCAG	TGGCTGCTGC	CAGTGGCGAT	AAGTCGTGTC	TTACCGGGTT	GGACTCAAGA	5700
	CGATAGTTAC	CGGATAAGGC	GCAGCGGTCG	GGCTGAACGG	GGGGTTCGTG	CACACAGCCC	5760
55	AGCTTGGAGC	GAACGACCTA	CACCGAACTO	G AGATACCTAC	AGCGTGAGCA	TTGAGAAAGC	5820
33	GCCACGCTTC	CCGAAGGGAG	AAAGGCGGAC	AGGTATCCGC	TAAGCGGCAG	GGTCGGAACA	5880
	GGAGAGCGCA	CGAGGGAGCT	TCCAGGGGG	AACGCCTGGT	r atctttatag	TCCTGTCGGG	5940
60						GCGGAGCCTA	6000
	TGGAAAAACG CCCTCAAGCC	CCAGCAACGO TCACTAAAA	C CGAGATGCGG GGTCCCTGCG	C CGCCTCGAG	r ACACCTGCGT r TACTAATCTG	CATGCTGAGA CCTTATTCTG	6060 6120
65	TTTTTGTTCC	CATGTTAAA	ATAGAGTAA	A TGCAGTATTO	C TCCACATAGA	GATATAGACT	6180

	TCTGAAATTC TAAGATTAGA ATTATTTACA AGAAGAAGTG GGGAA	6225-							
	(2) INFORMATION FOR SEQ ID NO:2:								
5	(i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 487 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear								
10	(ii) MOLECULE TYPE: DNA (genomic)								
15	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:								
	CCTCCCATCT AGAGGTTGTT CTCGGAACAC TCCTAAACTT TTCACCCCAA AACTCCTCAC	60							
20	CCTAAAGTTC GAAAAAACTG TTCCAAGAAC ATTTTTGAGA TAAAGGCCTC CTAGAACAAC	120							
	CTCAAAATGA CATTGCCAAA TGATAAGACA TGACTCCTTA GTTACGTAGG TTCCTTGATA	180							
25	GGACATGACT CCTTAGTTAC GTAGGTTCCT TGATAGGACA TGACTCCTTA GTTACGTAGA	240							
25	TTCCTTTGGT AGAACTCCCT AGTGATGTAA ACTTGTACTT TCCCTGCCCA GTTCTCCCCC	300							
	TTTGAGTTTT ACTATATAG CCTGTAAAAA ATTTTTGCTG ACCGTCGAGA CTCCTCTACC	360							
30	CTGTGCTAAG GTGTATGAGT TTCGACCCCA GAGCTCTGTG TGCTTCCATG TTGCTGCTTT	420							
	ATTTCGACCC CAGAGCTCTG GTCTGTGTGC TTTCATGTCG CTGCTTTATT AAATCTTGCC	480							
2.5	TTCTACA	487							
35	(2) INFORMATION FOR SEQ ID NO:3:								
40	 (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 366 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear 								
45	(ii) MOLECULE TYPE: DNA (genomic)								
50	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:								
50	CCTCCCATCT AGAAAACATT TTTGAGATAA AGGCTTCCTG GAACAACCTC AAAATGAACC	60							
	AGGTACTCCT TAGTTACGTA GGTTCCTTGA TAGGACATGA CTCCTTAGTT ACATAGATTC	120							
55	CTTTGGCAGA ACTCCCTAGT GATGTAAACT TGTACTTTCC CTGCCCAGTT CTCCCCCTTT	180							
	GAGTTTTACT ATATAAGCCT GTGAAAAATT TTGGCTGACC GTCGAGACTC CTCTACCCTG	240							
60	TGCTAAGGTG TATGAGTTTC GACCCCAGAG CTCTGTGTGC TTCCATGTTG CTGCTTTATT	300							
00	TCGACCCCAG AGCTCTGGTC TGTGTGCTTT CATGTTGCTG CCTTATTAAA TCTTGCCTTC TACATT	360 366							
65	(2) INFORMATION FOR SEQ ID NO:4: (i) SEQUENCE CHARACTERISTICS:								

	(A) LENGTH: 304 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear	-	
5	(ii) MOLECULE TYPE: DNA (genomic)		
10	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:		
	CCTCCCATCT AGAGATTGTT CCCAGAACAC TCCTGAACTC TTCACCCCAG AATGCATGCC	60	
15	TGAACTCCTC ACCCTAGAGT TCGAACCCTC CCAACTAAAG ACTGTTCCAA GAACATTTTT	120	
	GAGATAAGGG CCTCCTGGAA CAACCTCAGA ATGAACCGGG TACATTGCCA AATAATAGGA	180	
20	CATGACCCCT TAGTTACGTA AAATCCCTTG GCAGAACCCC TTGTCCCTTG GCAGAACCCC	240	
20	TTAGTTATGT AAACTTGTAC TTTCCCTACC CCGCTCTCCC CCCTTGAGTT TTTCCTATAT	300	
	AAGC	304	
25	(2) INFORMATION FOR SEQ ID NO:5:		
30	(i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 304 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear		
	(ii) MOLECULE TYPE: DNA (genomic)		
35			
	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:		
40	CCTCCCATCT AGAGAGTGTT CCCAGAACAC TCCTGAACTC TTCACCCCAG AATGCATTCC	60	
40	TGAACTCCTC ACCCTAGAGT TCGAACCCTC CCAACTAAAG ACTGTTCCAA GAACATTTTT	120	
	GAGATAAGGG CCTCCTGGAA CAACCTCAGA ATGAACCGGA TACATTGCCA AATAATAGGA	180	
45	CATGACCCCT TAGTTACGTA GAATCCCTTG GCAGAACCCC TTGTCCCTTG GCAGAACCCC	240	
	TTAGTTATGT AAACTTGTAC TTTCCCTACC CCGCTCTCCC CCCTTGAGTT TTTCCTATAT	300	
50	AAGC	304	
50	(2) INFORMATION FOR SEQ ID NO:6:		
55	(i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 304 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear		
60	(ii) MOLECULE TYPE: DNA (genomic)		
	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:		
65	CCTCCCATCT AGAGAGTGTT CCCAGAACAC TCCTGAACTC TTCACCCCAG AATGCATTCC	60	

	TGAACTCCTC ATCCTAGAGT TCGAACCCTC CCAACTAAAG ACTGTTCCAA GAACATTTTT	120						
_	GAGATAAGGG CCTCCTGGAA CAACCTCAGA ATGAACCTGG TACATTGCCA AATAATAGGA	180						
5	CATGACCCTT TAGTTACGTA GAATCCCTTG GCAGAACCCC TTGTCCCTTG GCAGAACCCC	240						
	TTAGTTATGC AAACTTGTAC TTTCTCTGCC CCGCTCTCCC CCCTTGAGTT TTTCCTATAT	300						
10	AAGC	304						
	(2) INFORMATION FOR SEQ ID NO:7:							
15	 (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 304 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear 							
20	(ii) MOLECULE TYPE: DNA (genomic)							
25	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:							
	CCTCCCATCT AGAGAGTGTT CCCAGAACAC TCCTGAACTC TTCACCTCAA AATGCATTCC	60						
	TGAACTCCTC ACCCTAGAGT TCGAACCCTC CCAACTAAAG ACTGTTCCAA GAACATTTTT	120						
30	GAGATAAGGG CCTCCTGGAA CAACCTCAGA ATGAACCAGG TACATTGCCA AATAATAGGA	180						
	CATGACCCTT TAGTTACGTA GAATCCCTTG GCAGAACCCC TTGTCCCTTG GCAGAACCCC	240						
35	TTAGTTATGC AAACTTGTAC TTTCTCTGCC CCGCTCTCCC CCCTTGAGTT TTTCCTATAT	300						
	AAGC	304						
40	(2) INFORMATION FOR SEQ ID NO:8:							
40	(i) SEQUENCE CHARACTERISTICS:							
	(A) LENGTH: 305 base pairs(B) TYPE: nucleic acid(C) STRANDEDNESS: single							
45	(D) TOPOLOGY: linear							
	(ii) MOLECULE TYPE: DNA (genomic)							
50								
	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:							
<i>.</i>	CCTCCCATCT AGAGATTGTT CCCAGAACAC TCCTGAACTC TTCACCCCAG AATGCATTCC	60						
55	TGAACTCCTC ACCCTAGAGT TCGAACCCTC CCAACTAAAG ACTGTTCCAA GAACATTTTT	120						
	GAGATAAGGG CCTCCTGGAA CAACCTCAGA ATGAACCGGA TACATTGCCA AATAATAGGA	180						
60	CATGACCCCT TAGTTACGTA GAATTCCCTT GGCAGAACCC CTTGTCCCTT GGCAGAACCC	240						
	CTTAGTTATG CAAACTTGTA CTTTCCCTGC CCCGCTCTCC CCCCTTGAGG TTTTCCTATA	300						
65	TAAGC 3							
,	(2) INFORMATION FOR SEQ ID NO:9:							

. 5	 (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 305 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear 	•
10	(ii) MOLECULE TYPE: DNA (genomic)	
10		
	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:9:	
15	CCTCCCATCT AGAGAGTGTT CCCAGAACAC TCCTGAACTC TTCACCCCAG AATGCATTCC	60
	TGAACCCCTC ACCCTAGAGT TCGAACCCTC CCAACTAAAG ACTGTTCCAA GAACATTTTT	120
20	GAGATAAGGG CCTCCTGGAA CAACCTCAGA ATGAACCAGG TACATTGCCA AATAATAGGA	180
20	CATGACCCCT TAGTTACGTA GAATTCCCTT GGCAGAACCC CTTGTCCCTT GGCAGAACCC	240
	CTTAGTTATG CGAACTTGTA CTTTCCCTGC CCCGCTCTCC CCCCTTGAGT TTTTCCTATA	300
25	TAAGC	305
	(2) INFORMATION FOR SEQ ID NO:10:	
30	 (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 306 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear 	
35	(ii) MOLECULE TYPE: DNA (genomic)	
40	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:10:	
	CCCTCCCATC TAGAGAGTGT TCCCAGAACA CTCCTGAACT CTTCATCCCA GAATGCATTC	60
45	CTGAACTCCT CACCCTATAG TTCGAACCCT CCCAACTAAA GACTGTTCCA AGAACATTTT	120
43	TGAGATAAGG GCCTCCTGGA ACAACCTCAG AATGAACCGG GTACATTGCC AAATAATAGG	180
	ACATGACCCC TTAGTTACGT AGAATTCCCT TGGCAGAACC CCTTGTCGCT TGGCAGAACC	240
50	CCTTAGTTAT GTAAACTTGT ACTTTCCCTG CCCCGCTCTC CCCCCTTGAG TTTTTACTAT	300
	ATAAGC	306
55	(2) INFORMATION FOR SEQ ID NO:11: (i) SEQUENCE CHARACTERISTICS:	
60	(A) LENGTH: 305 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear	
	(ii) MOLECULE TYPE: DNA (genomic)	

	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:11:	-
	CCTCCCATCT AGAGAGTGTT CCCAAAACAC TCCTGAACTC TTCACCCCAG AATGCATTCC	60
5	TGAACTCCTC ACCCTAAAGT TCAAACCCTC CCAACTAAAG ACTGTTCCAA GAACATTTTT	120
	GAGATAAGGG CCTCCTGGAA CAACCTCAGA ATGAACCGGG TACATTGCCA AATAATAGGA	180
10	CATGACCCCT TAGTTACACA GAATTCCCTT GGCAAAACCC CTTGTCCCTT GGCAGAACCC	240
10	CTTAGTTATG CAAACTTGTA CTTTCCCTGC CCAGCTCTCC CCCCTTGAGT TTTTCCTATA	300
	TAAGC	305
15	(2) INFORMATION FOR SEQ ID NO:12:	
20	 (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 304 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear 	
	(ii) MOLECULE TYPE: DNA (genomic)	
25		
	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:12:	
30	CCTCCCATCT AGAGAGTGTT CCCAGAACAC TCCTGAACTC TTCACCCCAG AATGCATTCC	60
50	TGAACTCCTC ACCCTAGAGT TTGAACCCTC CCAACTAAAG ACTGTTCCAA GAACATCTTT	120
	GAGATAAGGG CCTCCTGGAA CAACCTCAGA ATGAACCGGG TACATTGCCA AATAATAGGA	180
35	CATGACCCCT TAGTTACGTA GAATTCCCTT GGCAGAACCC CTTGTCGCTT GGCAGAACCC	240
	CTTAGTTATG CAAACTTGTA CTTTCCCTGC CCCGCTCTCC CCCTTGAGTT TTTCCTATAT	300
40	AAGC	304
	(2) INFORMATION FOR SEQ ID NO:13:	
45	(i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 303 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear	
50	(ii) MOLECULE TYPE: DNA (genomic)	
55	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:13:	
	CCTCCCATCT AGAGAGTGTT CCCAGAACAC TCCTAAACTC TTCACCCCAG AATGCATTCC TGAACTCCTC ACCCTAGAGT TCGAACCCTT CCAACTAAAG ACTGTTCCAA GAACATTTTT	60 120
60	GAGATAAGGG CCTCCTGGAA CAACCTCAAA ATGAACCGGG TACATTGCCA AATGATAGGA	180
	CATGACCCCT TAGTTACGTA GATTCCCTTG GCAGAACCCC TTGTCCCTTG GCAGAACCCC	240
65	CTAGTGATGT AAACTTGTAC TTTCCCTGCC CAGCTCTCCC CCCTTGAGTT TTCCTATATA	300
0,5	AGC	303

(2) INFORMATION FOR SEQ ID NO:14:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 8657 base pairs

(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic) 10

5

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:14: 15

	TGAAGAATAA	AAAATTACTG	GCCTCTTGTG	AGAACATGAA	CTTTCACCTC	GGAGCCCACC	60
20	CCCTCCCATC	TGGAAAACAT	ACTTGAGAAA	AACATTTTCT	GGAACAACCA	CAGAATGTTT	120
20	CAACAGGCCA	GATGTATTGC	CAAACACAGG	ATATGACTCT	TTGGTTGAGT	AAATTTGTGG	180
	TTGTTAAACT	TCCCCTATTC	CCTCCCCATT	CCCCCTCCCA	GTTTGTGGTT	TTTTCCTTTA	240
25	AAAGCTTGTG	AAAAATTTGA	GTCGTCGTCG	AGACTCCTCT	ACCCTGTGCA	AAGGTGTATG	300
	AGTTTCGACC	CCAGAGCTCT	GTGTGCTTTC	TGTTGCTGCT	TTATTTCGAC	CCCAGAGCTC	360
20	TGGTCTGTGT	GCTTTCATGT	CGCTGCTTTA	TTAAATCTTA	CCTTCTACAT	TTTATGTATG	420
30	GTCTCAGTGT	CTTCTTGGGT	ACGCGGCTGT	CCCGGGACTT	GAGTGTCTGA	GTGAGGGTCT	480
	TCCCTCGAGG	GTCTTTCATT	TGGTACATGG	GCCGGGAATT	CGAGAATCTT	TCATTTGGTG	540
35	CATTGGCCGG	GAATTCGAAA	ATCTTTCATT	TGGTGCATTG	GCCGGGAAAC	AGCGCGACCA	600
	CCCAGAGGTC	CTAGACCCAC	TTAGAGGTAA	GATTCTTTGT	TCTGTTTTGG	TCTGATGTCT	660
40	GTGTTCTGAT	GTCTGTGTTC	TGTTTCTAAG	TCTGGTGCGA	TCGCAGTTTC	AGTTTTGCGG	720
40	ACGCTCAGTG	AGACCGCGCT	CCGAGAGGGA	GTGCGGGGTG	GATAAGGATA	GACGTGTCCA	780
	GGTGTCCACC	GTCCGTTCGC	CCTGGGAGAC	GTCCCAGGAG	GAACAGGGGA	GGATCAGGGA	840
45	CGCCTGGTGG	ACCCCTTTGA	AGGCCAAGAG	ACCATTTGGG	GTTGCGAGAT	CGTGGGTTCG	900
	AGTCCCACCT	CGTGCCCAGT	TGCGAGATCG	TGGGTTCGAG	TCCCACCTCG	TGTTTTGTTG	960
50	CGAGATCGTG	GGTTCGAGTC	CCACCTCGCG	TCTGGTCACG	GGATCGTGGG	TTCGAGTCCC	1020
30	ACCTCGTGTT	TTGTTGCGAG	ATCGTGGGTT	CGAGTCCCAC	CTCGCGTCTG	GTCACGGGAT	1080
	CGTGGGTTCG	AGTCCCACCT	CGTGCAGAGG	GTCTCAATTG	GCCGGCCTTA	GAGAGGCCAT	1140
55	CTGATTCTTC	TGGTTTCTCT	TTTTGTCTTA	GTCTCGTGTC	CGCTCTTGTT	GTGACTACTG	1200
	TTTTTCTAAA	AATGGGACAA	TCTGTGTCCA	CTCCCCTTTC	TCTGACTCTG	GTTCTGTCGC	1260
60	TTGGTAATTT	TGTTTGTTTA	CGTTTGTTT	TGTGAGTCGT	CTATGTTGTC	TGTTACTATC	1320
60	TTGTTTTTGT	TTGTGGTTTA	CGGTTTCTGI	GTGTGTCTTG	TGTGTCTCTT	TGTGTTCAGA	1380
	CTTGGACTGA	TGACTGACGA	CTGTTTTAP	GTTATGCCTT	CTAAAATAAG	CCTAAAAATC	1440
65	CTGTCAGATC	: CCTATGCTGA	CCACTTCCTT	TCAGATCAAC	AGCTGCCCTT	ACGTATCGAT-	1500

			* C C C C * T T T C T	CCNACCECCA	CCCCCATCAA	ምምርርርርርርር	1560
					GCGGGATCAA		1620
•					GTGTGCGTTT		
5					CCGGAAACCT		1680
	TCTTGACGAG	CATTCCTAGG	GGTCTTTCCC	CTCTCGCCAA	AGGAATGCAA	GGTCTGTTGA	1740
10	ATGTCGTGAA	GGAAGCAGTT	CCTCTGGAAG	CTTCTTGAAG	ACAAACAACG	TCTGTAGCGA	1800
10	CCCTTTGCAG	GCAGCGGAAC	CCCCCACCTG	GCGACAGGTG	CCTCTGCGGC	CAAAAGCCAC	1860
	GTGTATAAGA	TACACCTGCA	AAGGCGGCAC	AACCCCAGTG	CCACGTTGTG	AGTTGGATAG	1920
15	TTGTGGAAAG	AGTCAAATGG	CTCTCCTCAA	GCGTATTCAA	CAAGGGGCTG	AAGGATGCCC	1980
	AGAAGGTACC	CCATTGTATG	GGATCTGATC	TGGGGCCTCG	GTGCACATGC	TTTACATGTG	2040
•	TTTAGTCGAG	GTTAAAAAAA	CGTCTAGGCC	CCCCGAACCA	CGGGGACGTG	GTTTTCCTTT	2100
20	GAAAAACACG	ATAATAATCA	TGGGCGCGGA	TCCCGTCGTT	TTACAACGTC	GTGACTGGGA	2160
	AAACCCTGGC	GTTACCCAAC	TTAATCGCCT	TGCAGCACAT	CCCCCTTTCG	CCAGCTGGCG	2220
25	TAATAGCGAA	GAGGCCCGCA	CCGATCGCCC	TTCCCAACAG	TTGCGCAGCC	TGAATGGCGA	2280
	ATGGCGCTTT	GCCTGGTTTC	CGGCACCAGA	AGCGGTGCCG	GAAAGCTGGC	TGGAGTGCGA	2340
20	TCTTCCTGAG	GCCGATACTG	TCGTCGTCCC	CTCAAACTGG	CAGATGCACG	GTTACGATGC	2400
30	GCCCATCTAC	ACCAACGTAA	CCTATCCCAT	TACGGTCAAT	CCGCCGTTTG	TTCCCACGGA	2460
	GAATCCGACG	GGTTGTTACT	CGCTCACATT	TAATGTTGAT	GAAAGCTGGC	TACAGGAAGG	2520
35	CCAGACGCGA	ATTATTTTTG	ATGGCGTTAA	CTCGGCGTTT	CATCTGTGGT	GCAACGGGCG	2580
	CTGGGTCGGT	TACGGCCAGG	ACAGTCGTTT	GCCGTCTGAA	TTTGACCTGA	GCGCATTTTT	2640
	ACGCGCCGGA	GAAAACCGCC	TCGCGGTGAT	GGTGCTGCGT	TGGAGTGACG	GCAGTTATCT	2700
40	GGAAGATCAG	GATATGTGGC	GGATGAGCGG	CATTTTCCGT	GACGTCTCGT	TGCTGCATAA	2760
	ACCGACTACA	CAAATCAGCG	ATTTCCATGT	TGCCACTCGC	TTTAATGATG	ATTTCAGCCG	2820
45	CGCTGTACTG	GAGGCTGAAG	TTCAGATGTG	CGGCGAGTTG	CGTGACTACC	TACGGGTAAC	2880
	AGTTTCTTTA	TGGCAGGGTG	AAACGCAGGT	CGCCAGCGGC	: ACCGCGCCTT	TCGGCGGTGA	2940
	AATTATCGAT	GAGCGTGGTG	GTTATGCCGA	TCGCGTCACA	CTACGTCTGA	ACGTCGAAAA	3000
50	CCCGAAACTG	TGGAGCGCCG	AAATCCCGAA	TCTCTATCGT	GCGGTGGTTG	AACTGCACAC	3060
	CGCCGACGGC	ACGCTGATTO	AAGCAGAAGC	: CTGCGATGTC	GGTTTCCGCG	AGGTGCGGAT	3120
55	TGAAAATGGT	CTGCTGCTGC	: TGAACGGCAA	GCCGTTGCTC	ATTCGAGGCG	TTAACCGTCA	3180
					ACGATGGTGC		3240
						ACCATCCGCT	3300
60						ATATTGAAAC	3360
	CCACGGCATO	GTGCCAATG	ATCGTCTGAC	CGATGATCC	GCGCTGGCTAC	CGGCGATGAG	3420
65	CGAACGCGT	A ACGCGAATGO	TGCAGCGCG	A TCGTAATCA	CCGAGTGTGA	TCATCTGGTC	3480
U.J	GCTGGGGAA	r GAATCAGGC	ACGGCGCTA	A TCACGACGC	CTGTATCGCT	GGATCAAATC	3540

							_
5	TGTCGATCCT	TCCCGCCCGG	TGCAGTATGA	AGGCGGCGGA	GCCGACACCA	CGGCCACCGA	3600
	TATTATTTGC	CCGATGTACG	CGCGCGTGGA	TGAAGACCAG	CCCTTCCCGG	CTGTGCCGAA	3660
	ATGGTCCATC	AAAAAATGGC	TTTCGCTACC	TGGAGAGACG	CGCCCGCTGA	TCCTTTGCGA	3720
	ATACGCCCAC	GCGATGGGTA	ACAGTCTTGG	CGGTTTCGCT	AAATACTGGC	AGGCGTTTCG	3780
10	TCAGTATCCC	CGTTTACAGG	GCGGCTTCGT	CTGGGACTGG	GTGGATCAGT	CGCTGATTAA	3840
	ATATGATGAA	AACGGCAACC	CGTGGTCGGC	TTACGGCGGT	GATTTTGGCG	ATACGCCGAA	3900
	CGATCGCCAG	TTCTGTATGA	ACGGTCTGGT	CTTTGCCGAC	CGCACGCCGC	ATCCAGCGCT	3960
15	GACGGAAGCA	AAACACCAGC	AGCAGTTTTT	CCAGTTCCGT	TTATCCGGGC	AAACCATCGA	4020
	AGTGACCAGC	GAATACCTGT	TCCGTCATAG	CGATAACGAG	CTCCTGCACT	GGATGGTGGC	4080
20	GCTGGATGGT	AAGCCGCTGG	CAAGCGGTGA	AGTGCCTCTG	GATGTCGCTC	CACAAGGTAA	4140
	ACAGTTGATT	GAACTGCCTG	AACTACCGCA	GCCGGAGAGC	GCCGGGCAAC	TCTGGCTCAC	4200
25	AGTACGCGTA	GTGCAACCGA	ACGCGACCGC	ATGGTCAGAA	GCCGGGCACA	TCAGCGCCTG	4260
25	GCAGCAGTGG	CGTCTGGCGG	AAAACCTCAG	TGTGACGCTC	CCCGCCGCGT	CCCACGCCAT	4320
	CCCGCATCTG	ACCACCAGCG	AAATGGATTT	TTGCATCGAG	CTGGGTAATA	AGCGTTGGCA	4380
30	ATTTAACCGC	CAGTCAGGCT	TTCTTTCACA	GATGTGGATT	GGCGATAAAA	AACAACTGCT	4440
	GACGCCGCTG	CGCGATCAGT	TCACCCGTGC	ACCGCTGGAT	AACGACATTG	GCGTAAGTGA	4500
25	AGCGACCCGC	ATTGACCCTA	ACGCCTGGGT	CGAACGCTGG	AAGGCGGCGG	GCCATTACCA	4560
35	GGCCGAAGCA	GCGTTGTTGC	AGTGCACGGC	AGATACACTT	GCTGATGCGG	TGCTGATTAC	4620
	GACCGCTCAC	GCGTGGCAGC	ATCAGGGGAA	AACCTTATTT	ATCAGCCGGA	AAACCTACCG	4680
40	GATTGATGGT	AGTGGTCAAA	TGGCGATTAC	CGTTGATGTT	GAAGTGGCGA	GCGATACACC	4740
	GCATCCGGCG	CGGATTGGCC	TGAACTGCCA	GCTGGCGCAG	GTAGCAGAGC	GGGTAAACTG	4800
45	GCTCGGATTA	GGGCCGCAAG	AAAACTATCC	CGACCGCCTT	ACTGCCGCCT	GTTTTGACCG	4860
43	CTGGGATCTG	CCATTGTCAG	ACATGTATAC	CCCGTACGTC	TTCCCGAGCG	AAAACGGTCT	4920
	GCGCTGCGGG	ACGCGCGAAT	TGAATTATGG	CCCACACCAG	TGGCGCGGCG	ACTTCCAGTT	4980
50	CAACATCAGC	CGCTACAGTC	AACAGCAACT	GATGGAAACC	AGCCATCGCC	ATCTGCTGCA	5040
	CGCGGAAGAA	GGCACATGGC	TGAATATCGA	CGGTTTCCAT	ATGGGGATTG	GTGGCGACGA	5100
55	CTCCTGGAGC GTTGGTCTGG	CCGTCAGTAT	CGGCGGAATT AATAATAACC	CCAGCTGAGC GGGCAGGGGG	GCCGGTCGCT GATCCGAAGG	ACCATTACCA CGGGGACAGC	5160 5220
	AGTGCAGTGG	TGGACAGAAA	GCAAGTGATC	TAGGCCAGCA	GCCTCCCTAA	AGGGACTTCA	5280
60	GCCCACAAAG	CCAAACTTGT	GGCTTTAATA	CAAGCTCTGT	: AAATGGTAAA	AAAAAAAAAG	5340
	TCTACACGGA	CAGCAGGTAT	GCTCTTGCCA	CTGTACAGAG	CAATATACAG	ACAAAGAGAA	5400
	CTGTTGACAT	CTGCAGAGAA	AGACCTAAGA	TGCTGTGGCT	r aaaagaaatc	AGATGGCAAA	5460
65	TCTAACCGCC	CAGGCATCCT	AAAGAGCAAT	GATCCTGAC	A GTCTGAAGAC	TATCAAGTTA	5520

	TAGACAAATT	AAGACTGGTA	ааааааассс	TGTATAAAAT	AGTAAAAACT	GAAAAAGAA	5580
	AACTAGTCCT	CTCATGAGAA	GACAGACCTG	ACATCTACTG	AAAAATAGAC	TTTACTGGAA	5640
5	AAAATATGTG	TATGAATACC	TTCTAGTTTT	TGTGAACGTT	CTCAAGATGG	ATAAAAGCTT	5700
	TTCCTTGTAA	AACGAGACTG	ATCAGATAGT	CATCAAGAAG	ATTGTTAAAG	AAAATTTTCC	5760
	AAGGTTCGGA	GTGCCAAAAG	CAATAGTGTC	AGATAATGGT	CCTGCCTTTG	TTGCCCAGGT	5820
10	AAGTCAGGGT	GTGGCCAAGT	ATTTAGAGGT	CAAATGAAAA	TTCCATTGTG	TGTACAGACC	5880
	TCAGAGCTCA	GGAAAGATAA	AAAAGAATAA	ATAAAACTCT	AAACAGACCT	TGACAAAATT	5940
15	AATCCTAGAG	ACTGGCACAG	ACTTACTTGG	TACTCCTTCC	CCTTGCCCTA	TTTAGAACTG	6000
	AGAATACTCC	CTCTTGATTC	GGTTTTACTC	TTTTTAAGAT	CCTTTATGGG	GCTCCTATGC	6060
20	CATCACTGTC	TTAAATGATG	TGTTTAAACC	TATGTTGTTA	TAATAATGAT	CTATATGTTA	6120
20	AGTTAAAAGG	CTTGCAGGTG	GTGCAGAAAG	AAGTCTGGTC	ACAACTGGCT	ACAGTGAACA	6180
	AGCTGGGTAC	CCCAAGGACA	TCTTACCAGT	TCCAGCCAGA	GATCTGATCT	ACGATCCCCG	6240
25	GGTCGACCCG	GGTCGACCCT	GTGGAATGTG	TGTCAGTTAG	GGTGTGGAAA	GTCCCCAGGC	6300
	TCCCCAGCAG	GCAGAAGTAT	GCAAAGCATG	CATCTCAATT	AGTCAGCAAC	CAGGTGTGGA	6360
20	AAGTCCCCAG	GCTCCCCAGC	AGGCAGAAGT	ATGCAAAGCA	TGCATCTCAA	TTAGTCAGCA	6420
30	ACCATAGTCC	CGCCCCTAAC	TCCGCCCATC	CCGCCCCTAA	CTCCGCCCAG	TTCCGCCCAT	6480
	TCTCCGCCCC	ATGGCTGACT	AATTTTTTT	ATTTATGCAG	AGGCCGAGGC	CGCCTCGGCC	6540
35	TCTGAGCTAT	TCCAGAAGTA	GTGAGGAGGC	TTTTTTGGAG	GCCTAGGCTT	TTGCAAAAAG	6600
	CTTCACGCTG	CCGCAAGCAC	TCAGGGCGCA	AGGGCTGCTA	AAGGAAGCGG	AACACGTAGA	6660
40	AAGCCAGTCC	GCAGAAACGG	TGCTGACCCC	GGATGAATGT	CAGCTACTGG	GCTATCTGGA	6720
40	CAAGGGAAAA	CGCAAGCGCA	AAGAGAAAGC	AGGTAGCTTG	CAGTGGGCTT	ACATGGCGAT	6780
	AGCTAGACTG	GGCGGTTTTA	TGGACAGCAA	GCGAACCGGA	ATTGCCAGCT	GGGGCGCCCT	6840
45	CTGGTAAGGT	TGGGAAGCCC	TGCAAAGTAA	ACTGGATGGC	TTTCTTGCCG	CCAAGGATCT	6900
	GATGGCGCAG	GGGATCAAGA	TCTGATCAAG	G AGACAGGATG	AGGATCGTTT	CGCATGATTG	6960
50	AACAAGATGG	ATTGCACGCA	GGTTCTCCGG	CCGCTTGGGT	GGAGAGGCTA	TTCGGCTATG	7020
50	ACTGGGCACA	ACAGACAATC	GGCTGCTCTC	ATGCCGCCGT	GTTCCGGCTG	TCAGCGCAGG	7080
	GGCGCCCGGT	TCTTTTTGTC	AAGACCGACG	TGTCCGGTGC	CCTGAATGAA TTGCGCAGCT	CTGCAGGACG	7140 7200
55					AGTGCCGGGG		7260
						ATGCGGCGGC	7320
60						CGCATCGAGC	7380
00						GAAGAGCATC	7440
65						GACGGCGAGG	7500
						AATGGCCGCT	7560
	0. 00. 00.						

		_				
	TTTCTGGATT CATCGACTGT GGCCGGCTGG GTGTGGCGGA CCGCTATCAG GACATAGCGT	7620				
5	TGGCTACCCG TGATATTGCT GAAGAGCTTG GCGGCGAATG GGCTGACCGC TTCCTCGTGC	7680				
	TTTACGGTAT CGCCGCTCCC GATTCGCAGC GCATCGCCTT CTATCGCCTT CTTGACGAGT	7740				
	TCTTCTGAGC GGGACTCTGG GGTTCGAAAT GACCGACCAA GCGACGCCCA ACCTGCCATC	7800				
10	ACGAGATTTC GATTCCACCG CCGCCTTCTA TGAAAGGTTG GGCTTCGGAA TCGTTTTCCG	7860				
	GGACGGAATT CGTAATCTGC TGCTTGCAAA CAAAAAAACC ACCGCTACCA GCGGTGGTTT	7920				
15	GTTTGCCGGA TCAAGAGCTA CCAACTCTTT TTCCGAAGGT AACTGGCTTC AGCAGAGCGC	7980				
	AGATACCAAA TACTGTCCTT CTAGTGTAGC CGTAGTTAGG CCACCACTTC AAGAACTCTG	8040				
	TAGCACCGCC TACATACCTC GCTCTGCTAA TCCTGTTACC AGTGGCTGCT GCCAGTGGCG	8100				
20	ATAAGTCGTG TCTTACCGGG TTGGACTCAA GACGATAGTT ACCGGATAAG GCGCAGCGGT	8160				
	CGGGCTGAAC GGGGGGTTCG TGCACACAGC CCAGCTTGGA GCGAACGACC TACACCGAAC	8220				
	TGAGATACCT ACAGCGTGAG CATTGAGAAA GCGCCACGCT TCCCGAAGGG AGAAAGGCGG	8280				
25	ACAGGTATCC GGTAAGCGGC AGGGTCGGAA CAGGAGAGCG CACGAGGGAG CTTCCAGGGG	8340				
	GAAACGCCTG GTATCTTTAT AGTCCTGTCG GGTTTCGCCA CCTCTGACTT GAGCGTCGAT	8400				
30	TTTTGTGATG CTCGTCAGGG GGGCGGAGCC TATGGAAAAA CGCCAGCAAC GCCGAGATGC	8460				
	GCCGCCTCGA GTACACCTGC GTCATGCTGA GACCCTCAAG CCTCACTAAA AGGGTCCCTG	8520				
25	CCTAGTTCTG TTTACTAATC TGCCTTATTC TGTTTTTGTT CCCATGTTAA AGATAGAGTA	8580				
35	AATGCAGTAT TCTCCACATA GAGATATAGA CTTCTGAAAT TCTAAGATTA GAATTATTTA	8640				
	CAAGAAGAAG TGGGGAA	8657				
40	(2) INFORMATION FOR SEQ ID NO:15:					
	(i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 6359 base pairs					
45	(A) LENGTH: 6559 base parts (B) TYPE: nucleic acid (C) STRANDEDNESS: single					
45	(D) TOPOLOGY: linear					
	(ii) MOLECULE TYPE: DNA (genomic)					
50						
	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:15:					
55	TGAAGAATAA AAAATTACTG GCCTCTTGTG AGAACATGAA CTTTCACCTC GGAGCCCACC	60				
<i>-</i>	CCCTCCCATC TGGAAAACAT ACTTGAGAAA AACATTTTCT GGAACAACCA CAGAATGTTT	120				
60	CAACAGGCCA GATGTATTGC CAAACACAGG ATATGACTCT TTGGTTGAGT AAATTTGTGG	180				
	TTGTTAAACT TCCCCTATTC CCTCCCCATT CCCCCTCCCA GTTTGTGGTT TTTTCCTTTA	240				
	AAAGCTTGTG AAAAATTTGA GTCGTCGTCG AGACTCCTCT ACCCTGTGCA AAGGTGTATG	300				

65 AGTTTCGACC CCAGAGCTCT GTGTGCTTTC TGTTGCTGCT TTATTTCGAC CCCAGAGCTC 360

	TGGTCTGTGT	GCTTTCATGT	CGCTGCTTTA	ттааатстта	CCTTCTACAT	TTTATGTATG	420
	GTCTCAGTGT	CTTCTTGGGT	ACGCGGCTGT	CCCGGGACTT	GAGTGTCTGA	GTGAGGGTCT	480
5	TCCCTCGAGG	GTCTTTCATT	TGGTACATGG	GCCGGGAATT	CGAGAATCTT	TCATTTGGTG	540
	CATTGGCCGG	GAATTCGAAA	ATCTTTCATT	TGGTGCATTG	GCCGGGAAAC	AGCGCGACCA	600
	CCCAGAGGTC	CTAGACCCAC	TTAGAGGTAA	GATTCTTTGT	TCTGTTTTGG	TCTGATGTCT	660
10	GTGTTCTGAT	GTCTGTGTTC	TGTTTCTAAG	TCTGGTGCGA	TCGCAGTTTC	AGTTTTGCGG	720
	ACGCTCAGTG	AGACCGCGCT	CCGAGAGGGA	GTGCGGGGTG	GATAAGGATA	GACGTGTCCA	780
15	GGTGTCCACC	GTCCGTTCGC	CCTGGGAGAC	GTCCCAGGAG	GAACAGGGGA	GGATCAGGGA	840
	CGCCTGGTGG	ACCCCTTTGA	AGGCCAAGAG	ACCATTTGGG	GTTGCGAGAT	CGTGGGTTCG	900
20	AGTCCCACCT	CGTGCCCAGT	TGCGAGATCG	TGGGTTCGAG	TCCCACCTCG	TGTTTTGTTG	960
20	CGAGATCGTG	GGTTCGAGTC	CCACCTCGCG	TCTGGTCACG	GGATCGTGGG	TTCGAGTCCC	1020
	ACCTCGTGTT	TTGTTGCGAG	ATCGTGGGTT	CGAGTCCCAC	CTCGCGTCTG	GTCACGGGAT	1080
25	CGTGGGTTCG	AGTCCCACCT	CGTGCAGAGG	GTCTCAATTG	GCCGGCCTTA	GAGAGGCCAT	1140
	CTGATTCTTC	TGGTTTCTCT	TTTTGTCTTA	GTCTCGTGTC	CGCTCTTGTT	GTGACTACTG	1200
20	TTTTTCTAAA	AATGGGACAA	TCTGTGTCCA	CTCCCCTTTC	TCTGACTCTG	GTTCTGTCGC	1260
30	TTGGTAATTT	TGTTTGTTTA	CGTTTGTTTT	TGTGAGTCGT	CTATGTTGTC	TGTTACTATC	1320
	TTGTTTTTGT	TTGTGGTTTA	CGGTTTCTGT	GTGTGTCTTG	TGTGTCTCTT	TGTGTTCAGA	1380
35	CTTGGACTGA	TGACTGACGA	CTGTTTTTAA	GTTATGCCTT	CTAAAATAAG	CCTAAAAATC	1440
	CTGTCAGATC	CCTATGCTGA	CCACTTCCTT	TCAGATCAAC	AGCTGCCCTT	ACTCGAGCTC	1500
40	AAGCTTCGAA	TTCTGCAGTC	GACGGTACCG	CGGCCGCTAA	CTAATAGCCC	ATTCTCCAAG	1560
40	GTACGTAGCG	GGGATCAATT	ccgcccccc	CCTAACGTTA	CTGGCCGAAG	CCGCTTGGAA	1620
	TAAGGCCGGT	GTGCGTTTGT	CTATATGTTA	TTTTCCACCA	TATTGCCGTC	TTTTGGCAAT	1680
45	GTGAGGGCCC	GGAAACCTGG	CCCTGTCTTC	TTGACGAGCA	TTCCTAGGGG	TCTTTCCCCT	1740
	CTCGCCAAAG	GAATGCAAGG	TCTGTTGAAT	GTCGTGAAGG	: AAGCAGTTCC	TCTGGAAGCT	1800
50	TCTTGAAGAC	AAACAACGTC	TGTAGCGACC	CTTTGCAGGC	: AGCGGAACCC	CCCACCTGGC	1860
50	GACAGGTGCC CCCCAGTGCC	TCTGCGGCCA ACGTTGTGAG	AAAGCCACGT	GTATAAGATA GTGGAAAGAG	CACCTGCAAA TCAAATGGCT	GGCGGCACAA CTCCTCAAGC	1920 1980
55	GTATTCAACA	AGGGGCTGAA	GGATGCCCAG	AAGGTACCCC	: ATTGTATGGG	ATCTGATCTG	2040
	GGGCCTCGGT	GCACATGCTT	TACATGTGTT	TAGTCGAGGT	тааааааас	TCTAGGCCCC	2100
	CCGAACCACG	GGGACGTGGT	TTTCCTTTGA	AAAACACGAT	ACGGGATCCA	CCGGTCGCCA	2160
60	CCATGGGTA	AGGAGAAGAA	CTTTTCACAG	GAGTTGTCCC	AATTCTTGTT	GAATTAGAŢG	2220
	GTGATGTTA	TGGGCACAA	A TTTTCTGTCA	GTGGAGAGG	G TGAAGGTGAT	GCAACATACG	2280
65	GAAAACTTAG	CCTTAAATTI	: ATTTGCACTA	CTGGAAAACT	r ACCTGTTCC	TGGCCAACAC	2340
	TTGTCACTAC	TTTCACTTA	GGTGTTCAAT	GCTTTTCAAG	G ATACCCAGAT	CATATGAAAC	2400

							-
	GGCATGACTT T	TTCAAGAGT	GCCATGCCCG	AAGGTTATGT	ACAGGAAAGA	ACTATATTTT	2460
	TCAAAGATGA C	GGGAACTAC	AAGACACGTG	CTGAAGTCAA	GTTTGAAGGT	GATACCCTTG	2520
5	TTAATAGAAT C	GAGTTAAAA	GGTATTGATT	TTAAAGAAGA	TGGAAACATT	CTTGGACACA	2580
	AATTGGAATA C	CAACTATAAC	TCACACAATG	TATACATCAT	GGCAGACAAA	CAAAAGAATG	2640
10	GAACCAAAGT T	CAACTTCAAA	ATTAGACACA	ACATTGAAGA	TGGAAGCGTT	CAACTAGCAG	2700
	ACCATTATCA A	ACAAAATACT	CCAATTGGCG	ATGGCCCTGT	CCTTTTACCA	GACAACCATT	2760
	ACCTGTCCAC A	ACAATCTGCC	CTTTCGAAAG	ATCCCAACGA	AAAGAGAGAC	CACATGGTCC	2820
15	TTCTTGAGTT T	TGTAACAGCT	GCTGGGATTA	CACATGGCAT	GGATGAAÇTA	TACAAGTCCG	2880
	GATCTAGATA A	ACTGTATCGA	TGGATCCGAA	GGCGGGGACA	GCAGTGCAGT	GGTGGACAGA	2940
20	AAGCAAGTGA T	CTAGGCCAG	CAGCCTCCCT	AAAGGGACTT	CAGCCCACAA	AGCCAAACTT	3000
	GTGGCTTTAA T	TACAAGCTCT	GTAAATGGTA	даааааааа	AGTCTACACG	GACAGCAGGT	3060
	ATGCTCTTGC (CACTGTACAG	AGCAATATAC	AGACAAAGAG	AACTGTTGAC	ATCTGCAGAG	3120
25	AAAGACCTAA (GATGCTGTGG	CTAAAAGAAA	TCAGATGGCA	AATCTAACCG	CCCAGGCATC	3180
	CTAAAGAGCA A	ATGATCCTGA	CAGTCTGAAG	ACTATCAAGT	TATAGACAAA	TTAAGACTGG	3240
30	TAAAAAAAAC (CCTGTATAAA	ATAGTAAAAA	CTGAAAAAAG	AAAACTAGTC	CTCTCATGAG	3300
	AAGACAGACC	TGACATCTAC	TGAAAAATAG	ACTTTACTGG	AAAAAATATG	TGTATGAATA	3360
	CCTTCTAGTT	TTTGTGAACG	TTCTCAAGAT	GGATAAAAGC	TTTTCCTTGT	AAAACGAGAC	3420
35	TGATCAGATA (GTCATCAAGA	AGATTGTTAA	AGAAAATTTT	CCAAGGTTCG	GAGTGCCAAA	3480
	AGCAATAGTG	TCAGATAATG	GTCCTGCCTT	TGTTGCCCAG	GTAAGTCAGG	GTGTGGCCAA	3540
40	GTATTTAGAG	GTCAAATGAA	AATTCCATTG	TGTGTACAGA	CCTCAGAGCT	CAGGAAAGAT	3600
	AAAAAAGAAT	ааатааааст	CTAAACAGAC	CTTGACAAAA	TTAATCCTAG	AGACTGGCAC	3660
	AGACTTACTT	GGTACTCCTT	CCCCTTGCCC	TATTTAGAAC	TGAGAATACT	CCCTCTTGAT	3720
45	TCGGTTTTAC	TCTTTTTAAG	ATCCTTTATO	GGGCTCCTAT	GCCATCACTG	TCTTAAATGA	3780
	TGTGTTTAAA	CCTATGTTGT	TATAATAAT	ATCTATATG	TAAGTTAAAA	GGCTTGCAGG	3840
50	TGGTGCAGAA	AGAAGTCTGG	TCACAACTGO	CTACAGTGA	CAAGCTGGGT	ACCCCAAGGA	3900 3960
					CGGGTCGACC		4020
					GCTCCCCAGC		4080
55						AGGCTCCCCA	4140
						CCCGCCCCTA	4200
60						CCATGGCTGA	4260
						ATTCCAGAAG	
						TGCCGCAAGC	4320
65	ACTCAGGGCG	CAAGGGCTGC	TAAAGGAAG	C GGAACACGT.	A GAAAGCCAG1	CCGCAGAAAC	4380

	GGTGCTGACC	CCGGATGAAT	GTCAGCTACT	GGGCTATCTG	GACAAGGGAA	AACGCAAGCG	4440
	CAAAGAGAAA	GCAGGTAGCT	TGCAGTGGGC	TTACATGGCG	ATAGCTAGAC	TGGGCGGTTT	4500
5	TATGGACAGC	AAGCGAACCG	GAATTGCCAG	CTGGGGCGCC	CTCTGGTAAG	GTTGGGAAGC	4560
	CCTGCAAAGT	AAACTGGATG	GCTTTCTTGC	CGCCAAGGAT	CTGATGGCGC	AGGGGATCAA	4620
10	GATCTGATCA	AGAGACAGGA	TGAGGATCGT	TTCGCATGAT	TGAACAAGAT	GGATTGCACG	4680
10	CAGGTTCTCC	GGCCGCTTGG	GTGGAGAGGC	TATTCGGCTA	TGACTGGGCA	CAACAGACAA	4740
	TCGGCTGCTC	TGATGCCGCC	GTGTTCCGGC	TGTCAGCGCA	GGGGCGCCCG	GTTCTTTTTG	4800
15	TCAAGACCGA	CCTGTCCGGT	GCCCTGAATG	AACTGCAGGA	CGAGGCAGCG	CGGCTATCGT	4860
	GGCTGGCCAC	GACGGGCGTT	CCTTGCGCAG	CTGTGCTCGA	CGTTGTCACT	GAAGCGGGAA	4920
	GGGACTGGCT	GCTATTGGGC	GAAGTGCCGG	GGCAGGATCT	CCTGTCATCT	CACCTTGCTC	4980
20	CTGCCGAGAA	AGTATCCATC	ATGGCTGATG	CAATGCGGCG	GCTGCATACG	CTTGATCCGG	5040
	CTACCTGCCC	ATTCGACCAC	CAAGCGAAAC	ATCGCATCGA	GCGAGCACGT	ACTCGGATGG	5100
25	AAGCCGGTCT	TGTCGATCAG	GATGATCTGG	ACGAAGAGCA	TCAGGGGCTC	GCGCCAGCCG	5160
	AACTGTTCGC	CAGGCTCAAG	GCGCGCATGC	CCGACGGCGA	GGATCTCGTC	GTGACCCATG	5220
20	GCGATGCCTG	CTTGCCGAAT	ATCATGGTGG	AAAATGGCCG	CTTTTCTGGA	TTCATCGACT	5280
30	GTGGCCGGCT	GGGTGTGGCG	GACCGCTATC	AGGACATAGC	GTTGGCTACC	CGTGATATTG	5340
	CTGAAGAGCT	TGGCGGCGAA	TGGGCTGACC	GCTTCCTCGT	GCTTTACGGT	ATCGCCGCTC	5400
35	CCGATTCGCA	GCGCATCGCC	TTCTATCGCC	TTCTTGACGA	GTTCTTCTGA	GCGGGACTCT	5460
	GGGGTTCGAA	ATGACCGACC	AAGCGACGCC	CAACCTGCCA	TCACGAGATT	TCGATTCCAC	5520
40	CGCCGCCTTC	TATGAAAGGT	TGGGCTTCGG	AATCGTTTTC	CGGGACGGAA	TTCGTAATCT	5580
40	GCTGCTTGCA	А ААСАААААА	CCACCGCTAC	CAGCGGTGGT	TTGTTTGCCG	GATCAAGAGC	5640
	TACCAACTCT	TTTTCCGAAG	GTAACTGGCT	TCAGCAGAGC	GCAGATACCA	AATACTGTCC	5700
45	TTCTAGTGTA	A GCCGTAGTTA	GGCCACCACT	TCAAGAACTO	TGTAGCACCG	CCTACATACC	5760
	TCGCTCTGCT	r AATCCTGTTA	CCAGTGGCTG	CTGCCAGTG	GCGATAAGTCG	TGTCTTACCG	5820
50	GGTTGGACTC	AAGACGATAG	TTACCGGATA	AGGCGCAGCG	GTCGGGCTGA A ACTGAGATAC	ACGGGGGGTT	5880 5940
50					GGACAGGTAI		6000
					G GGGAAACGCC		6060
55					G ATTTTTGTGA		6120
					r gcgccgcctc		6180
60						TGTTTACTAA	6240
60						ATTCTCCACA	6300
					T TACAAGAAG!		6359
65		MATION FOR					
	(Z) INFOR	PALION FOR	THE IN MOST	••			

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 6891 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

10

5

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:16:

	• •	_					
15	TGAAGAATAA	AAAATTACTG	GCCTCTTGTG	AGAACATGAA	CTTTCACCTC	GGAGCCCACC	60
	CCCTCCCATC	TGGAAAACAT	ACTTGAGAAA	AACATTTTCT	GGAACAACCA	CAGAATGTTT	120
20	CAACAGGCCA	GATGTATTGC	CAAACACAGG	ATATGACTCT	TTGGTTGAGT	AAATTTGTGG	180
20	TTGTTAAACT	TCCCCTATTC	CCTCCCCATT	CCCCTCCCA	GTTTGTGGTT	TTTTCCTTTA	240
	AAAGCTTGTG	AAAAATTTGA	GTCGTCGTCG	AGACTCCTCT	ACCCTGTGCA	AAGGTGTATG	300
25	AGTTTCGACC	CCAGAGCTCT	GTGTGCTTTC	TGTTGCTGCT	TTATTTCGAC	CCCAGAGCTC	360
	TGGTCTGTGT	GCTTTCATGT	CGCTGCTTTA	TTAAATCTTA	CCTTCTACAT	TTTATGTATG	420
30	GTCTCAGTGT	CTTCTTGGGT	ACGCGGCTGT	CCCGGGACTT	GAGTGTCTGA	GTGAGGGTCT	480
טכ	TCCCTCGAGG	GTCTTTCATT	TGGTACATGG	GCCGGGAATT	CGAGAATCTT	TCATTTGGTG	540
	CATTGGCCGG	GAATTCGAAA	ATCTTTCATT	TGGTGCATTG	GCCGGGAAAC	AGCGCGACCA	600
35	CCCAGAGGTC	CTAGACCCAC	TTAGAGGTAA	GATTCTTTGT	TCTGTTTTGG	TCTGATGTCT	660
	GTGTTCTGAT	GTCTGTGTTC	TGTTTCTAAG	TCTGGTGCGA	TCGCAGTTTC	AGTTTTGCGG	720
40	ACGCTCAGTG	AGACCGCGCT	CCGAGAGGGA	GTGCGGGGTG	GATAAGGATA	GACGTGTCCA	780
40	GGTGTCCACC	GTCCGTTCGC	CCTGGGAGAC	GTCCCAGGAG	GAACAGGGGA	GGATCAGGGA	840
	CGCCTGGTGG	ACCCCTTTGA	AGGCCAAGAG	ACCATTTGGG	GTTGCGAGAT	CGTGGGTTCG	900
45	AGTCCCACCT	CGTGCCCAGT	TGCGAGATCG	TGGGTTCGAG	TCCCACCTCG	TGTTTTGTTG	960
	CGAGATCGTG	GGTTCGAGTC	CCACCTCGCG	TCTGGTCACG	GGATCGTGGG	TTCGAGTCCC	1020
50	ACCTCGTGTT	TTGTTGCGAG	ATCGTGGGTT	CGAGTCCCAC	CTCGCGTCTG	GTCACGGGAT	1080
30	CGTGGGTTCG	AGTCCCACCT	CGTGCAGAGG	GTCTCAATTG	GCCGGCCTTA	GAGAGGCCAT	1140
	CTGATTCTTC	TGGTTTCTCT	TTTTGTCTTA	GTCTCGTGTC	CGCTCTTGTT	GTGACTACTG	1200
55	TTTTTCTAAA	AATGGGACAA	TCTGTGTCCA	CTCCCCTTTC	TCTGACTCTG	GTTCTGTCGC	1260
	TTGGTAATTT	TGTTTGTTTA	CGTTTGTTTT	TGTGAGTCGT	CTATGTTGTC	TGTTACTATC	1320
60	TTGTTTTTGT	TTGTGGTTTA	CGGTTTCTGT	GTGTGTCTTG	TGTGTCTCTT	TGTGTTCAGA	1380
00	CTTGGACTGA	TGACTGACGA	CTGTTTTTAA	GTTATGCCTT	CTAAAATAAG	CCTAAAAATC	1440
	CTGTCAGATC	CCTATGCTGA	CCACTTCCTT	TCAGATCAAC	AGCTGCCCTT	ACTCGAGCTC	1500
65	AAGCTTCGAA	TTCTGCAGTC	GACGGTACCG	CGGGGATCAA	TTCCGCCCCC	CCCCTAACGT	1560

	TACTGGCCGA	AGCCGCTTGG	AATAAGGCCG	GTGTGCGTTT	GTCTATATGT	TATTTTCCAC	1620
	CATATTGCCG	TCTTTTGGCA	ATGTGAGGGC	CCGGAAACCT	GGCCCTGTCT	TCTTGACGAG	1680
5	CATTCCTAGG	GGTCTTTCCC	CTCTCGCCAA	AGGAATGCAA	GGTCTGTTGA	ATGTCGTGAA	1740
	GGAAGCAGTT	CCTCTGGAAG	CTTCTTGAAG	ACAAACAACG	TCTGTAGCGA	CCCTTTGCAG	1800
10	GCAGCGGAAC	CCCCCACCTG	GCGACAGGTG	CCTCTGCGGC	CAAAAGCCAC	GTGTATAAGA	1860
10	TACACCTGCA	AAGGCGGCAC	AACCCCAGTG	CCACGTTGTG	AGTTGGATAG	TTGTGGAAAG	1920
	AGTCAAATGG	CTCTCCTCAA	GCGTATTCAA	CAAGGGGCTG	AAGGATGCCC	AGAAGGTACC	1980
15	CCATTGTATG	GGATCTGATC	TGGGGCCTCG	GTGCACATGC	TTTACATGTG	TTTAGTCGAG	2040
	GTTAAAAAAC	GTCTAGGCCC	CCCGAACCAC	GGGGACGTGG	TTTTCCTTTG	AAAAACACGA	2100
20	GCGGGATCAA	TTCCGCCCCC	CCCCTAACGT	TACTGGCCGA	AGCCGCTTGG	AATAAGGCCG	2160
20	GTGTGCGTTT	GTCTATATGT	TATTTTCCAC	CATATTGCCG	TCTTTTGGCA	ATGTGAGGGC	2220
	CCGGAAACCT	GGCCCTGTCT	TCTTGACGAG	CATTCCTAGG	GGTCTTTCCC	CTCTCGCCAA	2280
25	AGGAATGCAA	GGTCTGTTGA	ATGTCGTGAA	GGAAGCAGTT	CCTCTGGAAG	CTTCTTGAAG	2340
	ACAAACAACG	TCTGTAGCGA	CCCTTTGCAG	GCAGCGGAAC	CCCCCACCTG	GCGACAGGTG	2400
30	CCTCTGCGGC	CAAAAGCCAC	GTGTATAAGA	TACACCTGCA	AAGGCGGCAC	AACCCCAGTG	2460
30	CCACGTTGTG	AGTTGGATAG	TTGTGGAAAG	AGTCAAATGG	CTCTCCTCAA	GCGTATTCAA	2520
	CAAGGGGCTG	AAGGATGCCC	AGAAGGTACC	CCATTGTATG	GGATCTGATC	TGGGGCCTCG	2580
35	GTGCACATGC	TTTACATGTG	TTTAGTCGAG	GTTAAAAAAA	CGTCTAGGCC	CCCCGAACCA	2640
	CGGGGACGTG	GTTTTCCTTT	GAAAAACACG	ATACGGGATC	CACCGGTCGC	CACCATGGGT	2700
40	AAAGGAGAAG	AACTTTTCAC	AGGAGTTGTC	CCAATTCTTG	TTGAATTAGA	TGGTGATGTT	2760
40	AATGGGCACA	AATTTTCTGT	CAGTGGAGAG	GGTGAAGGTG	ATGCAACATA	CGGAAAACTT	2820
	ACCCTTAAAT	TTATTTGCAC	TACTGGAAAA	CTACCTGTTC	CATGGCCAAC	ACTTGTCACT	2880
45	ACTTTCACTT	ATGGTGTTCA	ATGCTTTTCA	AGATACCCAG	ATCATATGAA	ACGGCATGAC	2940
	TTTTTCAAGA	GTGCCATGCC	CGAAGGTTAT	GTACAGGAAA	GAACTATATT GTGATACCCT	TTTCAAAGAT TGTTAATAGA	3000 3060
50						CAAATTGGAA	3120
50						TGGAACCAAA	3180
					TTCAACTAGC		3240
55					CAGACAACCA		3300
					ACCACATGGT		3360
60						CGGATCTAGA	3420
					GTGGTGGACA		3480
						TTGTGGCTTT	3540
65	•				i	GTATGCTCTT	3600

	GCCACTGTAC 2	AGAGCAATAT	ACAGACAAAG	AGAACTGTTG	ACATCTGCAG	AGAAAGACCT	3660
_	AAGATGCTGT	GGCTAAAAGA	AATCAGATGG	CAAATCTAAC	CGCCCAGGCA	TCCTAAAGAG	3720
5	CAATGATCCT	GACAGTCTGA	AGACTATCAA	GTTATAGACA	AATTAAGACT	GGTAAAAAAA	3780
	ACCCTGTATA	AAATAGTAAA	AACTGAAAAA	AGAAAACTAG	TCCTCTCATG	AGAAGACAGA	3840
10	CCTGACATCT .	ACTGAAAAAT	AGACTTTACT	GGAAAAAATA	TGTGTATGAA	TACCTTCTAG	3900
	TTTTTGTGAA	CGTTCTCAAG	ATGGATAAAA	GCTTTTCCTT	GTAAAACGAG	ACTGATCAGA	3960
	TAGTCATCAA	GAAGATTGTT	AAAGAAAATT	TTCCAAGGTT	CGGAGTGCCA	AAAGCAATAG	4020
15	TGTCAGATAA	TGGTCCTGCC	TTTGTTGCCC	AGGTAAGTCA	GGGTGTGGCC	AAGTATTTAG	4080
	AGGTCAAATG	AAAATTCCAT	TGTGTGTACA	GACCTCAGAG	CTCAGGAAAG	ATAAAAAAGA	4140
20	АТАААТАААА	CTCTAAACAG	ACCTTGACAA	AATTAATCCT	AGAGACTGGC	ACAGACTTAC	4200
	TTGGTACTCC	TTCCCCTTGC	CCTATTTAGA	ACTGAGAATA	CTCCCTCTTG	ATTCGGTTTT	4260
0.5	ACTCTTTTTA	AGATCCTTTA	TGGGGCTCCT	ATGCCATCAC	TGTCTTAAAT	GATGTGTTTA	4320
25	AACCTATGTT	GTTATAATAA	TGATCTATAT	GTTAAGTTAA	AAGGCTTGCA	GGTGGTGCAG	4380
	AAAGAAGTCT	GGTCACAACT	GGCTACAGTG	AACAAGCTGG	GTACCCCAAG	GACATCTTAC	4440
30	CAGTTCCAGC	CAGAGATCTG	ATCTACGATC	CCCGGGTCGA	CCCGGGTCGA	CCCTGTGGAA	4500
	TGTGTGTCAG	TTAGGGTGTG	GAAAGTCCCC	AGGCTCCCCA	GCAGGCAGAA	GTATGCAAAG	4560
26	CATGCATCTC	AATTAGTCAG	CAACCAGGTG	TGGAAAGTCC	CCAGGCTCCC	CAGCAGGCAG	4620
35	AAGTATGCAA	AGCATGCATC	TCAATTAGTC	AGCAACCATA	GTCCCGCCCC	TAACTCCGCC	4680
	CATCCCGCCC	CTAACTCCGC	CCAGTTCCGC	CCATTCTCCG	CCCCATGGCT	GACTAATTTT	4740
40	TTTTATTTAT	GCAGAGGCCG	AGGCCGCCTC	GGCCTCTGAG	CTATTCCAGA	AGTAGTGAGG	4800
	AGGCTTTTTT	GGAGGCCTAG	GCTTTTGCAA	AAAGCTTCAC	GCTGCCGCAA	GCACTCAGGG	4860
15	CGCAAGGGCT	GCTAAAGGAA	GCGGAACACG	TAGAAAGCCA	GTCCGCAGAA	ACGGTGCTGA	4920
45	CCCCGGATGA AAGCAGGTAG	ATGTCAGCTA CTTGCAGTGG	CTGGGCTATC GCTTACATGG	TGGACAAGGG CGATAGCTAG	AAAACGCAAG ACTGGGCGGT	CGCAAAGAGA TTTATGGACA	4980 5040
50	GCAAGCGAAC	CGGAATTGCC	AGCTGGGGCG	CCCTCTGGTA	AGGTTGGGAA	GCCCTGCAAA	5100
50	GTAAACTGGA	TGGCTTTCTT	GCCGCCAAGG	ATCTGATGGC	GCAGGGGATC	AAGATCTGAT	5160
	CAAGAGACAG	GATGAGGATC	GTTTCGCATG	ATTGAACAAG	ATGGATTGCA	CGCAGGTTCT	5220
55	CCGGCCGCTT	GGGTGGAGAG	GCTATTCGGC	: TATGACTGG	CACAACAGAC	AATCGGCTGC	5280
	TCTGATGCCG	CCGTGTTCCG	GCTGTCAGCG	CAGGGGCGCC	CGGTTCTTT	TGTCAAGACC	5340
60	GACCTGTCCG	GTGCCCTGAA	TGAACTGCAG	GACGAGGCAG	G CGCGGCTATC	GTGGCTGGCC	5400
60	ACGACGGGCG	TTCCTTGCGC	AGCTGTGCTC	GACGTTGTC	A CTGAAGCGGG	AAGGGACTGG	5460
	CTGCTATTGG	GCGAAGTGCC	GGGGCAGGAT	CTCCTGTCA	r ctcaccttgo	TCCTGCCGAG	5520
65	AAAGTATCCA	TCATGGCTGA	TGCAATGCGG	G CGGCTGCAT	A CGCTTGATCO	GGCTACCTGC	5580

	CCATTCGACC	ACCAAGCGAA	ACATCGCATC	GAGCGAGCAC	GTACTCGGAT	GGAAGCCGGT	5640
	CTTGTCGATC	AGGATGATCT	GGACGAAGAG	CATCAGGGGC	TCGCGCCAGC	CGAACTGTTC	5700
5	GCCAGGCTCA	AGGCGCGCAT	GCCCGACGGC	GAGGATCTCG	TCGTGACCCA	TGGCGATGCC	5760
	TGCTTGCCGA	ATATCATGGT	GGAAAATGGC	CGCTTTTCTG	GATTCATCGA	CTGTGGCCGG	5820
10	CTGGGTGTGG	CGGACCGCTA	TCAGGACATA	GCGTTGGCTA	CCCGTGATAT	TGCTGAAGAG	5880
10	CTTGGCGGCG	AATGGGCTGA	CCGCTTCCTC	GTGCTTTACG	GTATCGCCGC	TCCCGATTCG	5940
	CAGCGCATCG	CCTTCTATCG	CCTTCTTGAC	GAGTTCTTCT	GAGCGGGACT	CTGGGGTTCG	6000
15	AAATGACCGA	CCAAGCGACG	CCCAACCTGC	CATCACGAGA	TTTCGATTCC	ACCGCCGCCT	6060
	TCTATGAAAG	GTTGGGCTTC	GGAATCGTTT	TCCGGGACGG	AATTCGTAAT	CTGCTGCTTG	6120
20	САААСААААА	AACCACCGCT	ACCAGCGGTG	GTTTGTTTGC	CGGATCAAGA	GCTACCAACT	6180
20	CTTTTTCCGA	AGGTAACTGG	CTTCAGCAGA	GCGCAGATAC	CAAATACTGT	CCTTCTAGTG	6240
	TAGCCGTAGT	TAGGCCACCA	CTTCAAGAAC	TCTGTAGCAC	CGCCTACATA	CCTCGCTCTG	6300
25	CTAATCCTGT	TACCAGTGGC	TGCTGCCAGT	GGCGATAAGT	CGTGTCTTAC	CGGGTTGGAC	6360
	TCAAGACGAT	AGTTACCGGA	TAAGGCGCAG	CGGTCGGGCT	GAACGGGGGG	TTCGTGCACA	6420
20	CAGCCCAGCT	TGGAGCGAAC	GACCTACACC	GAACTGAGAT	ACCTACAGCG	TGAGCATTGA	6480
30	GAAAGCGCCA	CGCTTCCCGA	AGGGAGAAAG	GCGGACAGGT	ATCCGGTAAG	CGGCAGGGTC	6540
	GGAACAGGAG	AGCGCACGAG	GGAGCTTCCA	GGGGGAAACG	CCTGGTATCT	TTATAGTCCT	6600
35	GTCGGGTTTC	GCCACCTCTG	ACTTGAGCGT	CGATTTTTGT	GATGCTCGTC	AGGGGGGCGG	6660
	AGCCTATGGA	AAAACGCCAG	CAACGCCGAG	ATGCGCCGCC	TCGAGTACAC	CTGCGTCATG	6720
40	CTGAGACCCT	CAAGCCTCAC	TAAAAGGGTC	CCTGCCTAGT	TCTGTTTACT	AATCTGCCTT	6780
40	ATTCTGTTTT	TGTTCCCATG	TTAAAGATAG	AGTAAATGCA	GTATTCTCCA	CATAGAGATA	6840
	TAGACTTCTG	AAATTCTAAG	ATTAGAATTA	TTTACAAGAA	GAAGTGGGGA	Α	6891
4550	(2) INFORM (i) S	(B) TYPE: n	RACTERISTIC 6321 base ucleic acid DNESS: sing	S: pairs			
	(ii) M	OLECULE TYP	E: DNA (gen	omic)			•
55							
		SEQUENCE DES					
60						GGAGCCCACC	60
						CAGAATGTTT	120
						AAATTTGTGG	180
65	TTGTTAAACT	TCCCCTATTC	CCTCCCCATT	CCCCCTCCCA	GTTTGTGGTI	TTTTCCTTTA	240

	AAAGCTTGTG	AAAAATTTGA	GTCGTCGTCG	AGACTCCTCT	ACCCTGTGCA	AAGGTGTATG	300
	AGTTTCGACC	CCAGAGCTCT	GTGTGCTTTC	TGTTGCTGCT	TTATTTCGAC	CCCAGAGCTC	360
5	TGGTCTGTGT	GCTTTCATGT	CGCTGCTTTA	ттааатстта	CCTTCTACAT	TTTATGTATG	420
	GTCTCAGTGT	CTTCTTGGGT	ACGCGGCTGT	CCCGGGACTT	GAGTGTCTGA	GTGAGGGTCT	480
	TCCCTCGAGG	GTCTTTCATT	TGGTACATGG	GCCGGGAATT	CGAGAATCTT	TCATTTGGTG	540
10	CATTGGCCGG	GAATTCGAAA	ATCTTTCATT	TGGTGCATTG	GCCGGGAAAC	AGCGCGACCA	600
	CCCAGAGGTC	CTAGACCCAC	TTAGAGGTAA	GATTCTTTGT	TCTGTTTTGG	TCTGATGTCT	660
15	GTGTTCTGAT	GTCTGTGTTC	TGTTTCTAAG	TCTGGTGCGA	TCGCAGTTTC	AGTTTTGCGG	720
	ACGCTCAGTG	AGACCGCGCT	CCGAGAGGGA	GTGCGGGGTG	GATAAGGATA	GACGTGTCCA	780
••	GGTGTCCACC	GTCCGTTCGC	CCTGGGAGAC	GTCCCAGGAG	GAACAGGGGA	GGATCAGGGA	840
20	CGCCTGGTGG	ACCCCTTTGA	AGGCCAAGAG	ACCATTTGGG	GTTGCGAGAT	CGTGGGTTCG	900
	AGTCCCACCT	CGTGCCCAGT	TGCGAGATCG	TGGGTTCGAG	TCCCACCTCG	TGTTTTGTTG	960
25	CGAGATCGTG	GGTTCGAGTC	CCACCTCGCG	TCTGGTCACG	GGATCGTGGG	TTCGAGTCCC	1020
	ACCTCGTGTT	TTGTTGCGAG	ATCGTGGGTT	CGAGTCCCAC	CTCGCGTCTG	GTCACGGGAT	1080
20	CGTGGGTTCG	AGTCCCACCT	CGTGCAGAGG	GTCTCAATTG	GCCGGCCTTA	GAGAGGCCAT	1140
30	CTGATTCTTC	TGGTTTCTCT	TTTTGTCTTA	GTCTCGTGTC	CGCTCTTGTT	GTGACTACTG	1200
	TTTTTCTAAA	AATGGGACAA	TCTGTGTCCA	стсссстттс	TCTGACTCTG	GTTCTGTCGC	1260
35	TTGGTAATTT	TGTTTGTTTA	CGTTTGTTTT	TGTGAGTCGT	CTATGTTGTC	TGTTACTATC	1320
	TTGTTTTTGT	TTGTGGTTTA	CGGTTTCTGT	GTGTGTCTTG	TGTGTCTCTT	TGTGTTCAGA	1380
40	CTTGGACTGA	TGACTGACGA	CTGTTTTTAA	GTTATGCCTT	CTAAAATAAG	CCTAAAAATC	1440
40	CTGTCAGATC	CCTATGCTGA	CCACTTCCTT	TCAGATCAAC	AGCTGCCCTT	ACTCGAGCTC	1500
	AAGCTTCGAA	TTCTGCAGTC	GACGGTACCG	CGGGGATCAA	TTCCGCCCCC	CCCCTAACGT	1560
45	TACTGGCCGA	AGCCGCTTGG	AATAAGGCCG	GTGTGCGTTT	GTCTATATGT	TATTTTCCAC	1620
	CATATTGCCG	TCTTTTGGCA	ATGTGAGGGC	CCGGAAACCT	GGCCCTGTCT	TCTTGACGAG	1680
50	CATTCCTAGG	GGTCTTTCCC	CTCTCGCCAA	AGGAATGCAA	GGTCTGTTGA	ATGTCGTGAA	1740
30	GGAAGCAGTT	CCTCTGGAAG	CTTCTTGAAG	ACAAACAACG	TCTGTAGCGA	CCCTTTGCAG	1800
	GCAGCGGAAC	CCCCCACCTG	GCGACAGGTG	CCTCTGCGGC	CAAAAGCCAC	GTGTATAAGA	1860
55	TACACCTGCA	AAGGCGGCAC	AACCCCAGTG	CCACGTTGTC	AGTTGGATAG	TTGTGGAAAG	1920
	AGTCAAATGG	CTCTCCTCAA	GCGTATTCAA	CAAGGGGCTC	AAGGATGCCC	AGAAGGTACC	1980
60	CCATTGTATG	GGATCTGATC	: TGGGGCCTCG	GTGCACATG	TTTACATGTG	TTTAGTCGAG	2040
60	GTTAAAAAAA	CGTCTAGGCC	CCCCGAACCA	CGGGGACGT	G GTTTTCCTTT	GAAAAACACG	2100
	ATACGGGATC	CACCGGTCGC	CACCATGGGT	· AAAGGAGAA	AACTTTTCAC	AGGAGTTGTC	2160
65	CCAATTCTTG	TTGAATTAGA	TGGTGATGT1	AATGGGCAC	A AATTTTCTGT	CAGTGGAGAG	2220

	GGTGAAGGTG	ATGCAACATA	CGGAAAACTT	ACCCTTAAAT	TTATTTGCAC	TACTGGAAAA	2280
	CTACCTGTTC	CATGGCCAAC	ACTTGTCACT	ACTTTCACTT	ATGGTGTTCA	ATGCTTTTCA	2340
. 5	AGATACCCAG	ATCATATGAA	ACGGCATGAC	TTTTTCAAGA	GTGCCATGCC	CGAAGGTTAT	2400
	GTACAGGAAA	GAACTATATT	TTTCAAAGAT	GACGGGAACT	ACAAGACACG	TGCTGAAGTC	2460
	AAGTTTGAAG	GTGATACCCT	TGTTAATAGA	ATCGAGTTAA	AAGGTATTGA	TTTTAAAGAA	2520
10	GATGGAAACA	TTCTTGGACA	CAAATTGGAA	TACAACTATA	ACTCACACAA	TGTATACATC	2580
	ATGGCAGACA	AACAAAAGAA	TGGAACCAAA	GTTAACTTCA	AAATTAGACA	CAACATTGAA	2640
15	GATGGAAGCG	TTCAACTAGC	AGACCATTAT	СААСААААТА	CTCCAATTGG	CGATGGCCCT	2700
	GTCCTTTTAC	CAGACAACCA	TTACCTGTCC	ACACAATCTG	CCCTTTCGAA	AGATCCCAAC	2760
	GAAAAGAGAG	ACCACATGGT	CCTTCTTGAG	TTTGTAACAG	CTGCTGGGAT	TACACATGGC	2820
20	ATGGATGAAC	TATACAAGTC	CGGATCTAGA	TAACTGTATC	GATGGATCCG	AAGGCGGGGA	2880
	CAGCAGTGCA	GTGGTGGACA	GAAAGCAAGT	GATCTAGGCC	AGCAGCCTCC	CTAAAGGGAC	2940
25	TTCAGCCCAC	AAAGCCAAAC	TTGTGGCTTT	AATACAAGCT	CTGTAAATGG	TAAAAAAAA	3000
	AAAGTCTACA	CGGACAGCAG	GTATGCTCTT	GCCACTGTAC	AGAGCAATAT	ACAGACAAAG	3060
20	AGAACTGTTG	ACATCTGCAG	AGAAAGACCT	AAGATGCTGT	GGCTAAAAGA	AATCAGATGG	3120
30	CAAATCTAAC	CGCCCAGGCA	TCCTAAAGAG	CAATGATCCT	GACAGTCTGA	AGACTATCAA	3180
	GTTATAGACA	AATTAAGACT	GGTAAAAAAA	. ACCCTGTATA	AAATAGTAAA	AACTGAAAAA	3240
35	AGAAAACTAG	TCCTCTCATG	AGAAGACAGA	CCTGACATCT	ACTGAAAAAT	AGACTTTACT	3300
	GGAAAAAATA	TGTGTATGAA	TACCTTCTAG	TTTTTGTGAA	CGTTCTCAAG	ATGGATAAAA	3360
40	GCTTTTCCTT	GTAAAACGAG	ACTGATCAGA	TAGTCATCAA	GAAGATTGTT	AAAGAAAATT	3420
40	TTCCAAGGTT	CGGAGTGCCA	AAAGCAATAG	TGTCAGATAA	TGGTCCTGCC	TTTGTTGCCC	3480
	AGGTAAGTCA	GGGTGTGGCC	AAGTATTTAG	AGGTCAAATG	AAAATTCCAT	TGTGTGTACA	3540 3600
45					CTCTAAACAG		3660
					TTCCCCTTGC		3720
50					AGATCCTTTA GTTATAATAA		3780
50					GITATAATAA GGTCACAACT		3840
					CAGAGATCTG		3900
55					TTAGGGTGTG		3960
						CAACCAGGTG	4020
CO						: TCAATTAGTC	4080
60						CCAGTTCCGC	4140
						AGGCCGCCTC	4200
65							4260
	GGCCTCTGAC	CTATTCCAGA	AGTAGTGAG	G AGGUTTTTT.	GONGGCC IAC	GCTTTTGCAA	.200

	AAAGCTTCAC	GCTGCCGCAA	GCACTCAGGG	CGCAAGGGCT	GCTAAAGGAA	GCGGAACACG	4320
5	TAGAAAGCCA	GTCCGCAGAA	ACGGTGCTGA	CCCCGGATGA	ATGTCAGCTA	CTGGGCTATC	4380
3	TGGACAAGGG	AAAACGCAAG	CGCAAAGAGA	AAGCAGGTAG	CTTGCAGTGG	GCTTACATGG	4440
	CGATAGCTAG	ACTGGGCGGT	TTTATGGACA	GCAAGCGAAC	CGGAATTGCC	AGCTGGGGCG	4500
10	CCCTCTGGTA	AGGTTGGGAA	GCCCTGCAAA	GTAAACTGGA	TGGCTTTCTT	GCCGCCAAGG	4560
	ATCTGATGGC	GCAGGGGATC	AAGATCTGAT	CAAGAGACAG	GATGAGGATC	GTTTCGCATG	4620
15	ATTGAACAAG	ATGGATTGCA	CGCAGGTTCT	CCGGCCGCTT	GGGTGGAGAG	GCTATTCGGC	4680
13	TATGACTGGG	CACAACAGAC	AATCGGCTGC	TCTGATGCCG	CCGTGTTCCG	GCTGTCAGCG	4740
	CAGGGGCGCC	CGGTTCTTTT	TGTCAAGACC	GACCTGTCCG	GTGCCCTGAA	TGAACTGCAG	4800
20	GACGAGGCAG	CGCGGCTATC	GTGGCTGGCC	ACGACGGGCG	TTCCTTGCGC	AGCTGTGCTC	4860
	GACGTTGTCA	CTGAAGCGGG	AAGGGACTGG	CTGCTATTGG	GCGAAGTGCC	GGGGCAGGAT	4920
25	CTCCTGTCAT	CTCACCTTGC	TCCTGCCGAG	AAAGTATCCA	TCATGGCTGA	TGCAATGCGG	4980
23	CGGCTGCATA	CGCTTGATCC	GGCTACCTGC	CCATTCGACC	ACCAAGCGAA	ACATCGCATC	5040
	GAGCGAGCAC	GTACTCGGAT	GGAAGCCGGT	CTTGTCGATC	AGGATGATCT	GGACGAAGAG	5100
30	CATCAGGGGC	TCGCGCCAGC	CGAACTGTTC	GCCAGGCTCA	AGGCGCGCAT	GCCCGACGGC	5160
	GAGGATCTCG	TCGTGACCCA	TGGCGATGCC	TGCTTGCCGA	ATATCATGGT	GGAAAATGGC	5220
35	CGCTTTTCTG	GATTCATCGA	CTGTGGCCGG	CTGGGTGTGG	CGGACCGCTA	TCAGGACATA	5280
33	GCGTTGGCTA	CCCGTGATAT	TGCTGAAGAG	CTTGGCGGCG	AATGGGCTGA	CCGCTTCCTC	5340
	GTGCTTTACG	GTATCGCCGC	TCCCGATTCG	CAGCGCATCG	CCTTCTATCG	CCTTCTTGAC	5400
40	GAGTTCTTCT	GAGCGGGACT	CTGGGGTTCG	AAATGACCGA	CCAAGCGACG	CCCAACCTGC	5460
	CATCACGAGA TCCGGGACGG	TTTCGATTCC AATTCGTAAT	ACCGCCGCCT CTGCTGCTTG	TCTATGAAAG CAAACAAAAA	GTTGGGCTTC AACCACCGCT	GGAATCGTTT ACCAGCGGTG	5520 5580
45	GTTTGTTTGC	CGGATCAAGA	GCTACCAACT	CTTTTTCCGA	AGGTAACTGG	CTTCAGCAGA	5640
	GCGCAGATAC	CAAATACTGT	CCTTCTAGTG	TAGCCGTAGT	TAGGCCACCA	CTTCAAGAAC	5700
5 0	TCTGTAGCAC	CGCCTACATA	CCTCGCTCTG	CTAATCCTGT	TACCAGTGGC	TGCTGCCAGT	5760
50	GGCGATAAGT	CGTGTCTTAC	CGGGTTGGAC	TCAAGACGAT	AGTTACCGGA	TAAGGCGCAG	5820
	CGGTCGGGCT	GAACGGGGG	TTCGTGCACA	CAGCCCAGCT	TGGAGCGAAC	GACCTACACC	5880
55	GAACTGAGAT	ACCTACAGCG	TGAGCATTGA	GAAAGCGCCA	CGCTTCCCGA	AGGGAGAAAG	5940
	GCGGACAGGT	ATCCGGTAAG	CGGCAGGGTC	GGAACAGGAG	AGCGCACGAG	GGAGCTTCCA	6000
60	GGGGGAAACG	CCTGGTATCT	TTATAGTCCT	GTCGGGTTTC	GCCACCTCTG	ACTTGAGCGT	6060
60	CGATTTTTGT	GATGCTCGTC	AGGGGGGCGG	AGCCTATGGA	AAAACGCCAG	CAACGCCGAG	6120
	ATGCGCCGCC	TCGAGTACAC	CTGCGTCAT	CTGAGACCCT	CAAGCCTCAC	TAAAAGGGTC	6180
65	CCTGCCTAGT	TCTGTTTACT	AATCTGCCTT	ATTCTGTTT	TGTTCCCATC	TTAAAGATAG	6240

	AGTAAATGCA GTATTCTCCA CATAGAGATA TAGACTTCTG AAATTCTAAG ATTAGAATTA	6300								
	TTTACAAGAA GAAGTGGGGA A	6321								
5	(2) INFORMATION FOR SEQ ID NO:18:									
10	(i) SEQUENCE CHARACTERISTICS:(A) LENGTH: 5754 base pairs(B) TYPE: nucleic acid									
10	<pre>(C) STRANDEDNESS: single (D) TOPOLOGY: linear</pre>									
	(ii) MOLECULE TYPE: DNA (genomic)									
15										
	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:18:									
20	TGAAGAATAA AAAATTACTG GCCTCTTGTG AGAACATGAA CTTTCACCTC GGAGCCCACC	60								
20	CCCTCCCATC TGGAAAACAT ACTTGAGAAA AACATTTTCT GGAACAACCA CAGAATGTTT	120								
	CAACAGGCCA GATGTATTGC CAAACACAGG ATATGACTCT TTGGTTGAGT AAATTTGTGG	180								
25	TTGTTAAACT TCCCCTATTC CCTCCCCATT CCCCCTCCCA GTTTGTGGTT TTTTCCTTTA	240								
		300								
20	AAAGCTTGTG AAAAATTTGA GTCGTCGTCG AGACTCCTCT ACCCTGTGCA AAGGTGTATG	360								
30	AGTTTCGACC CCAGAGCTCT GTGTGCTTTC TGTTGCTGCT TTATTTCGAC CCCAGAGCTC									
	TGGTCTGTGT GCTTTCATGT CGCTGCTTTA TTAAATCTTA CCTTCTACAT TTTATGTATG	420								
35	GTCTCAGTGT CTTCTTGGGT ACGCGGCTGT CCCGGGACTT GAGTGTCTGA GTGAGGGTCT	480								
	TCCCTCGAGG GTCTTTCATT TGGTACATGG GCCGGGAATT CGAGAATCTT TCATTTGGTG	540								
	CATTGGCCGG GAATTCGAAA ATCTTTCATT TGGTGCATTG GCCGGGAAAC AGCGCGACCA	600								
40	CCCAGAGGTC CTAGACCCAC TTAGAGGTAA GATTCTTTGT TCTGTTTTGG TCTGATGTCT	660								
	GTGTTCTGAT GTCTGTTCT TGTTTCTAAG TCTGGTGCGA TCGCAGTTTC AGTTTTGCGG	720								
45	ACGCTCAGTG AGACCGCGCT CCGAGAGGGA GTGCGGGGTG GATAAGGATA GACGTGTCCA	780								
	GGTGTCCACC GTCCGTTCGC CCTGGGAGAC GTCCCAGGAG GAACAGGGGA GGATCAGGGA	840								
	CGCCTGGTGG ACCCCTTTGA AGGCCAAGAG ACCATTTGGG GTTGCGAGAT CGTGGGTTCG	900								
50	AGTCCCACCT CGTGCCCAGT TGCGAGATCG TGGGTTCGAG TCCCACCTCG TGTTTTGTTG	960								
	CGAGATCGTG GGTTCGAGTC CCACCTCGCG TCTGGTCACG GGATCGTGGG TTCGAGTCCC	1020								
55	ACCTCGTGTT TTGTTGCGAG ATCGTGGGTT CGAGTCCCAC CTCGCGTCTG GTCACGGGAT	1080								
23	CGTGGGTTCG AGTCCCACCT CGTGCAGAGG GTCTCAATTG GCCGGCCTTA GAGAGGCCAT	1140								
	CTGATTCTTC TGGTTTCTCT TTTTGTCTTA GTCTCGTGTC CGCTCTTGTT GTGACTACTG	1200								
60	TTTTTCTAAA AATGGGACAA TCTGTGTCCA CTCCCCTTTC TCTGACTCTG GTTCTGTCGC	1260								
	TTGGTAATTT TGTTTGTTTA CGTTTGTTTT TGTGAGTCGT CTATGTTGTC TGTTACTATC	1320								
65	TTGTTTTTGT TTGTGGTTTA CGGTTTCTGT GTGTGTCTTG TGTGTCTCTT TGTGTTCAGA	1380								
65		1440								

CTTGGACTGA TGACTGACGA CTGTTTTTAA GTTATGCCTT CTAAAATAAG CCTAAAAATC 1440

	CTGTCAGATC	CCTATGCTGA	CCACTTCCTT	TCAGATCAAC	AGCTGCCCTT	ACTCGAGCTC	1500
	AAGCTTCGAA	TTCTGCAGTC	GACGGTACCG	cggcccgg	ATCCACCGGT	CGCCACCATG	1560
5	GGTAAAGGAG	AAGAACTTTT	CACAGGAGTT	GTCCCAATTC	TTGTTGAATT	AGATGGTGAT	1620
	GTTAATGGGC	ACAAATTTTC	TGTCAGTGGA	GAGGGTGAAG	GTGATGCAAC	ATACGGAAAA	1680
10	CTTACCCTTA	AATTTATTTG	CACTACTGGA	AAACTACCTG	TTCCATGGCC	AACACTTGTC	1740
	ACTACTTTCA	CTTATGGTGT	TCAATGCTTT	TCAAGATACC	CAGATCATAT	GAAACGGCAT	1800
	GACTTTTTCA	AGAGTGCCAT	GCCCGAAGGT	TATGTACAGG	AAAGAACTAT	ATTTTTCAAA	1860
15	GATGACGGGA	ACTACAAGAC	ACGTGCTGAA	GTCAAGTTTG	AAGGTGATAC	CCTTGTTAAT	1920
	AGAATCGAGT	TAAAAGGTAT	TGATTTTAAA	GAAGATGGAA	ACATTCTTGG	ACACAAATTG	1980
20	GAATACAACT	ATAACTCACA	CAATGTATAC	ATCATGGCAG	ACAAACAAAA	GAATGGAACC	2040
	AAAGTTAACT	TCAAAATTAG	ACACAACATT	GAAGATGGAA	GCGTTCAACT	AGCAGACCAT	2100
25	TATCAACAAA	ATACTCCAAT	TGGCGATGGC	CCTGTCCTTT	TACCAGACAA	CCATTACCTG	2160
25	TCCACACAAT	CTGCCCTTTC	GAAAGATCCC	AACGAAAAGA	GAGACCACAT	GGTCCTTCTT	2220
	GAGTTTGTAA	CAGCTGCTGG	GATTACACAT	GGCATGGATG	AACTATACAA	GTCCGGATCT	2280
30	AGATAACTGT	ATCGATGGAT	CCGAAGGCGG	GGACAGCAGT	GCAGTGGTGG	ACAGAAAGCA	2340
	AGTGATCTAG	GCCAGCAGCC	TCCCTAAAGG	GACTTCAGCC	CACAAAGCCA	AACTTGTGGC	2400
25	TTTAATACAA	GCTCTGTAAA	TGGTAAAAA	AAAAAAGTCT	ACACGGACAG	CAGGTATGCT	2460
35	CTTGCCACTG	TACAGAGCAA	TATACAGACA	AAGAGAACTG	TTGACATCTG	CAGAGAAAGA	2520
	CCTAAGATGC	TGTGGCTAAA	AGAAATCAGA	TGGCAAATCT	AACCGCCCAG	GCATCCTAAA	2580
40	GAGCAATGAT AAAACCCTGT	CCTGACAGTC ATAAAATAGT	TGAAGACTAT AAAAACTGAA	CAAGTTATAG AAAAGAAAAC	ACAAATTAAG TAGTCCTCTC	ACTGGTAAAA ATGAGAAGAC	2640 2700
	AGACCTGACA	TCTACTGAAA	AATAGACTTT	ACTGGAAAAA	ATATGTGTAT	GAATACCTTC	2760
45	TAGTTTTTGT	GAACGTTCTC	AAGATGGATA	AAAGCTTTTC	CTTGTAAAAC	GAGACTGATC	2820
	AGATAGTCAT	CAAGAAGATT	GTTAAAGAAA	ATTTTCCAAG	GTTCGGAGTG	CCAAAAGCAA	2880
50	TAGTGTCAGA	TAATGGTCCT	GCCTTTGTTG	CCCAGGTAAG	TCAGGGTGTG	GCCAAGTATT	2940
30	TAGAGGTCAA	ATGAAAATTC	CATTGTGTGT	ACAGACCTCA	GAGCTCAGGA	AAGATAAAAA	3000
	AGAATAAATA	AAACTCTAAA	CAGACCTTGA	СААААТТААТ	CCTAGAGACT	GGCACAGACT	3060
55	TACTTGGTAC	TCCTTCCCCT	TGCCCTATT	AGAACTGAGA	ATACTCCCTC	TTGATTCGGT	3120
	TTTACTCTTT	TTAAGATCCI	TTATGGGGCT	CCTATGCCAT	CACTGTCTT	AATGATGTGT	3180
60	TTAAACCTAT	GTTGTTATAA	TAATGATCT	A TATGTTAAG1	TAAAAGGCTT	GCAGGTGGTG	3240
υu	CAGAAAGAAG	TCTGGTCAC	ACTGGCTACA	A GTGAACAAGO	TGGGTACCCC	AAGGACATCT	3300
	TACCAGTTCC	AGCCAGAGAT	CTGATCTAC	ATCCCCGGGT	CGACCCGGG	CGACCCTGTG	3360
65	GAATGTGTGT	CAGTTAGGG	GTGGAAAGT	CCCAGGCTCC	CCAGCAGGC	A GAAGTATGCA	3420

	AAGCATGCAT	CTCAATTAGT	CAGCAACCAG	GTGTGGAAAG	TCCCCAGGCT	CCCCAGCAGG	3480
	CAGAAGTATG	CAAAGCATGC	ATCTCAATTA	GTCAGCAACC	ATAGTCCCGC	CCCTAACTCC	3540
5	GCCCATCCCG	CCCCTAACTC	CGCCCAGTTC	CGCCCATTCT	CCGCCCCATG	GCTGACTAAT	3600
	TTTTTTTTTT	TATGCAGAGG	CCGAGGCCGC	CTCGGCCTCT	GAGCTATTCC	AGAAGTAGTG	3660
10	AGGAGGCTTT	TTTGGAGGCC	TAGGCTTTTG	CAAAAAGCTT	CACGCTGCCG	CAAGCACTCA	3720
10	GGGCGCAAGG	GCTGCTAAAG	GAAGCGGAAC	ACGTAGAAAG	CCAGTCCGCA	GAAACGGTGC	3780
	TGACCCCGGA	TGAATGTCAG	CTACTGGGCT	ATCTGGACAA	GGGAAAACGC	AAGCGCAAAG	3840
15	AGAAAGCAGG	TAGCTTGCAG	TGGGCTTACA	TGGCGATAGC	TAGACTGGGC	GGTTTTATGG	3900
	ACAGCAAGCG	AACCGGAATT	GCCAGCTGGG	GCGCCCTCTG	GTAAGGTTGG	GAAGCCCTGC	3960
20	AAAGTAAACT	GGATGGCTTT	CTTGCCGCCA	AGGATCTGAT	GGCGCAGGGG	ATCAAGATCT	4020
20	GATCAAGAGA	CAGGATGAGG	ATCGTTTCGC	ATGATTGAAC	AAGATGGATT	GCACGCAGGT	4080
	TCTCCGGCCG	CTTGGGTGGA	GAGGCTATTC	GGCTATGACT	GGGCACAACA	GACAATCGGC	4140
25	TGCTCTGATG	CCGCCGTGTT	CCGGCTGTCA	GCGCAGGGGC	GCCCGGTTCT	TTTTGTCAAG	4200
	ACCGACCTGT	CCGGTGCCCT	GAATGAACTG	CAGGACGAGG	CAGCGCGGCT	ATCGTGGCTG	4260
30	GCCACGACGG	GCGTTCCTTG	CGCAGCTGTG	CTCGACGTTG	TCACTGAAGC	GGGAAGGGAC	4320
30	TGGCTGCTAT	TGGGCGAAGT	GCCGGGGCAG	GATCTCCTGT	CATCTCACCT	TGCTCCTGCC	4380
	GAGAAAGTAT	CCATCATGGC	TGATGCAATG	CGGCGGCTGC	ATACGCTTGA	TCCGGCTACC	4440
35	TGCCCATTCG	ACCACCAAGC	GAAACATCGC	ATCGAGCGAG	CACGTACTCG	GATGGAAGCC	4500
	GGTCTTGTCG	ATCAGGATGA	TCTGGACGAA	GAGCATCAGG	GGCTCGCGCC	AGCCGAACTG	4560
40	TTCGCCAGGC GCCTGCTTGC	TCAAGGCGCG CGAATATCAT	CATGCCCGAC GGTGGAAAAT	GGCGAGGATC GGCCGCTTTT	TCGTCGTGAC CTGGATTCAT	CCATGGCGAT CGACTGTGGC	4620 4680
	CGGCTGGGTG	TGGCGGACCG	CTATCAGGAC	ATAGCGTTGG	CTACCCGTGA	TATTGCTGAA	4740
	GAGCTTGGCG	GCGAATGGGC	TGACCGCTTC	CTCGTGCTTT	ACGGTATCGC	CGCTCCCGAT	4800
45	TCGCAGCGCA	TCGCCTTCTA	TCGCCTTCTT	GACGAGTTCT	TCTGAGCGGG	ACTCTGGGGT	4860
	TCGAAATGAC	CGACCAAGCG	ACGCCCAACC	TGCCATCACG	AGATTTCGAT	TCCACCGCCG	4920
50	CCTTCTATGA	AAGGTTGGGC	TTCGGAATCG	TTTTCCGGGA	CGGAATTCGT	AATCTGCTGC	4980
	TTGCAAACAA	AAAAACCACC	GCTACCAGCG	GTGGTTTGTT	TGCCGGATCA	AGAGCTACCA	5040
<i>c c</i>	ACTCTTTTTC	CGAAGGTAAC	TGGCTTCAGC	AGAGCGCAGA	TACCAAATAC	TGTCCTTCTA	5100
55	GTGTAGCCGT	AGTTAGGCCA	CCACTTCAAG	AACTCTGTAG	CACCGCCTAC	ATACCTCGCT	5160
	CTGCTAATCC	TGTTACCAGT	GGCTGCTGCC	: AGTGGCGATA	AGTCGTGTCT	TACCGGGTTG	5220
60	GACTCAAGAC	GATAGTTACC	GGATAAGGCG	CAGCGGTCGG	GCTGAACGGG	GGGTTCGTGC	5280
	ACACAGCCCA	GCTTGGAGCG	AACGACCTAC	ACCGAACTG	A GATACCTACA	GCGTGAGCAT	5340
65	TGAGAAAGCG	CCACGCTTCC	CGAAGGGAGA	AAGGCGGAC	GGTATCCGGT	AAGCGGCAGG	5400
υS	GTCGGAACAG	GAGAGCGCAC	GAGGGAGCTT	CCAGGGGGA	A ACGCCTGGTA	TCTTTATAGT	5460

		-
	CCTGTCGGGT TTCGCCACCT CTGACTTGAG CGTCGATTTT TGTGATGCTC GTCAGGGGGG	5520
5	CGGAGCCTAT GGAAAAACGC CAGCAACGCC GAGATGCGCC GCCTCGAGTA CACCTGCGTC	5580
J	ATGCTGAGAC CCTCAAGCCT CACTAAAAGG GTCCCTGCCT AGTTCTGTTT ACTAATCTGC	5640
	CTTATTCTGT TTTTGTTCCC ATGTTAAAGA TAGAGTAAAT GCAGTATTCT CCACATAGAG	5700
10	ATATAGACTT CTGAAATTCT AAGATTAGAA TTATTTACAA GAAGAAGTGG GGAA	5754
	(2) INFORMATION FOR SEQ ID NO:19:	
15	 (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 5754 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear 	
20	(ii) MOLECULE TYPE: DNA (genomic)	
,		
0.5		
25	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:19:	60
	TGAAGAATAA AAAATTACTG GCCTCTTGTG AGAACATGAA CTTTCACCTC GGAGCCCACC CCCTCCCATC TGGAAAACAT ACTTGAGAAA AACATTTTCT GGAACAACCA CAGAATGTTT	120
30	CAACAGGCCA GATGTATTGC CAAACACAGG ATATGACTCT TTGGTTGAGT AAATTTGTGG	180
	TTGTTAAACT TCCCCTATTC CCTCCCATT CCCCCTCCCA GTTTGTGGTT TTTTCCTTTA	240
2.5		300
35	AAAGCTTGTG AAAAATTTGA GTCGTCGTCG AGACTCCTCT ACCCTGTGCA AAGGTGTATG AGTTTCGACC CCAGAGCTCT GTGTGCTTTC TGTTGCTGCT TTATTTCGAC CCCAGAGCTC	360
	TGGTCTGTGT GCTTTCATGT CGCTGCTTTA TTAAATCTTA CCTTCTACAT TTTATGTATG	420
40	GTCTCAGTGT CTTCTTGGGT ACGCGGCTGT CCCGGGACTT GAGTGTCTGA GTGAGGGTCT	480
	TCCCTCGAGG GTCTTTCATT TGGTACATGG GCCGGGAATT CGAGAATCTT TCATTTGGTG	540
45	CATTGGCCGG GAATTCGAAA ATCTTTCATT TGGTGCATTG GCCGGGAAAC AGCGCGACCA	600
43	CCCAGAGGTC CTAGACCCAC TTAGAGGTAA GATTCTTTGT TCTGTTTTGG TCTGATGTCT	660
	GTGTTCTGAT GTCTGTGTTC TGTTTCTAAG TCTGGTGCGA TCGCAGTTTC AGTTTTGCGG	720
50	ACGCTCAGTG AGACCGCGCT CCGAGAGGGA GTGCGGGGTG GATAAGGATA GACGTGTCCA	780
	GGTGTCCACC GTCCGTTCGC CCTGGGAGAC GTCCCAGGAG GAACAGGGGA GGATCAGGGA	840
55	CGCCTGGTGG ACCCCTTTGA AGGCCAAGAG ACCATTTGGG GTTGCGAGAT CGTGGGTTCG	900
33	AGTCCCACCT CGTGCCCAGT TGCGAGATCG TGGGTTCGAG TCCCACCTCG TGTTTTGTTG	960
	CGAGATCGTG GGTTCGAGTC CCACCTCGCG TCTGGTCACG GGATCGTGGG TTCGAGTCCC	1020
60	ACCTCGTGTT TTGTTGCGAG ATCGTGGGTT CGAGTCCCAC CTCGCGTCTG GTCACGGGAT	1080
	CGTGGGTTCG AGTCCCACCT CGTGCAGAGG GTCTCAATTG GCCGGCCTTA GAGAGGCCAT	1140
	Coloration Wallocker Calachana attickering account to allahanasani	

65 CTGATTCTTC TGGTTTCTCT TTTTGTCTTA GTCTCGTGTC CGCTCTTGTT GTGACTACTG 1200

	TTTTTCTAAA	AATGGGACAA	TCTGTGTCCA	CTCCCCTTTC	TCTGACTCTG	GTTCTGTCGC	1260
	TTGGTAATTT	TGTTTGTTTA	CGTTTGTTTT	TGTGAGTCGT	CTATGTTGTC	TGTTACTATC	1320
5	TTGTTTTTGT	TTGTGGTTTA	CGGTTTCTGT	GTGTGTCTTG	TGTGTCTCTT	TGTGTTCAGA	1380
	CTTGGACTGA	TGACTGACGA	CTGTTTTTAA	GTTATGCCTT	СТААААТААС	ССТАААААТС	1440
10	CTGTCAGATC	CCTATGCTGA	CCACTTCCTT	TCAGATCAAC	AGCTGCCCTT	ACTCGAGCTC	1500
10	AAGCTTCGAA	TTCTGCAGTC	GACGGTACCG	CGGGCCCGGG	ATCCACCGGT	CGCCACCATG	1560
	GGTAAAGGAG	AAGAACTTTT	CACTGGAGTT	GTCCCAATTC	TTGTTGAATT	AGATGGTGAT	1620
15	GTTAATGGGC	ACAAATTTTC	TGTCAGTGGA	GAGGGTGAAG	GTGATGCAAC,	ATACGGAAAA	1680
	CTTACCCTTA	AATTTATTTG	CACTACTGGA	AAACTACCTG	TTCCATGGCC	AACACTTGTC	1740
20	ACTACTTTCT	CTTATGGTGT	TCAATGCTTT	TCAAGATACC	CAGATCATAT	GAAACGGCAT	1800
20	GACTTTTTCA	AGAGTGCCAT	GCCCGAAGGT	TATGTACAGG	AAAGAACTAT	ATTTTTCAAA	1860
	GATGACGGGA	ACTACAAGAC	ACGTGCTGAA	GTCAAGTTTG	AAGGTGATAC	CCTTGTTAAT	1920
25	AGAATCGAGT	TAAAAGGTAT	TGATTTTAAA	GAAGATGGAA	ACATTCTTGG	ACACAAATTG	1980
	GAATACAACT	ATAACTCACA	CAATGTATAC	ATCATGGCAG	ACAAACAAAA	GAATGGAACC	2040
30	AAAGTTAACT	TCAAAATTAG	ACACAACATT	GAAGATGGAA	GCGTTCAACT	AGCAGACCAT	2100
30	TATCAACAAA	ATACTCCAAT	TGGCGATGGC	CCTGTCCTTT	TACCAGACAA	CCATTACCTG	2160
	TCCACACAAT	CTGCCCTTTC	GAAAGATCCC	AACGAAAAGA	GAGACCACAT	GGTCCTTCTT	2220
35	GAGTTTGTAA	CAGCTGCTGG	GATTACACAT	GGCATGGATG	AACTATACAA	GTCCGGATCT	2280
	AGATAACTGT AGTGATCTAG	ATCGATGGAT GCCAGCAGCC	CCGAAGGCGG TCCCTAAAGG	GGACAGCAGT GACTTCAGCC	GCAGTGGTGG CACAAAGCCA	ACAGAAAGCA AACTTGTGGC	2340 2400
40	TTTAATACAA	GCTCTGTAAA	TGGTAAAAAA	AAAAAGTCT	ACACGGACAG	CAGGTATGCT	2460
	CTTGCCACTG	TACAGAGCAA	TATACAGACA	AAGAGAACTG	TTGACATCTG	CAGAGAAAGA	2520
45	CCTAAGATGC	TGTGGCTAAA	AGAAATCAGA	TGGCAAATCT	AACCGCCCAG	GCATCCTAAA	2580
40	GAGCAATGAT	CCTGACAGTC	TGAAGACTAT	CAAGTTATAG	ACAAATTAAG	ACTGGTAAAA	2640
	AAAACCCTGT	ATAAAATAGT	AAAAACTGAA	AAAAGAAAAC	TAGTCCTCTC	ATGAGAAGAC	2700
50	AGACCTGACA	TCTACTGAAA	AATAGACTTT	ACTGGAAAAA	ATATGTGTAT	GAATACCTTC	2760
	TAGTTTTTGT	GAACGTTCTC	AAGATGGATA	AAAGCTTTTC	CTTGTAAAAC	GAGACTGATC	2820
55	AGATAGTCAT	CAAGAAGATT	GTTAAAGAAA	ATTTTCCAAG	GTTCGGAGTG	CCAAAAGCAA	2880
33	TAGTGTCAGA	TAATGGTCCT	GCCTTTGTTG	CCCAGGTAAG	TCAGGGTGTG	GCCAAGTATT	2940
	TAGAGGTCAA	ATGAAAATTC	CATTGTGTGT	ACAGACCTCA	GAGCTCAGGA	AAGATAAAAA	3000
60	AGAATAAATA	AAACTCTAAA	CAGACCTTGA	A CAAAATTAAT	CCTAGAGACI	GGCACAGACT	3060
	TACTTGGTAC	TCCTTCCCCT	TGCCCTATT	AGAACTGAGA	ATACTCCCTC	TTGATTCGGT	3120
65	TTTACTCTTT	TTAAGATCCT	TTATGGGGC	CCTATGCCAT	CACTGTCTTA	AATGATGTGT	3180
00	TTAAACCTAT	GTTGTTATAA	TAATGATCT	A TATGTTAAGT	TAAAAGGCTT	GCAGGTGGTG	3240

	CAGAAAGAAG	TCTGGTCACA	ACTGGCTACA	GTGAACAAGC	TGGGTACCCC	AAGGACATCT	3300
_	TACCAGTTCC	AGCCAGAGAT	CTGATCTACG	ATCCCCGGGT	CGACCCGGGT	CGACCCTGTG	3360
5	GAATGTGTGT	CAGTTAGGGT	GTGGAAAGTC	CCCAGGCTCC	CCAGCAGGCA	GAAGTATGCA	3420
	AAGCATGCAT	CTCAATTAGT	CAGCAACCAG	GTGTGGAAAG	TCCCCAGGCT	CCCCAGCAGG	3480
10	CAGAAGTATG	CAAAGCATGC	ATCTCAATTA	GTCAGCAACC	ATAGTCCCGC	CCCTAACTCC	3540
	GCCCATCCCG	CCCCTAACTC	CGCCCAGTTC	CGCCCATTCT	CCGCCCCATG	GCTGACTAAT	3600
	TTTTTTTATT	TATGCAGAGG	CCGAGGCCGC	CTCGGCCTCT	GAGCTATTCC	AGAAGTAGTG	3660
15	AGGAGGCTTT	TTTGGAGGCC	TAGGCTTTTG	CAAAAAGCTT	CACGCTGCCG	CAAGCACTCA	3720
	GGGCGCAAGG	GCTGCTAAAG	GAAGCGGAAC	ACGTAGAAAG	CCAGTCCGCA	GAAACGGTGC	3780
20	TGACCCCGGA	TGAATGTCAG	CTACTGGGCT	ATCTGGACAA	GGGAAAACGC	AAGCGCAAAG	3840
	AGAAAGCAGG	TAGCTTGCAG	TGGGCTTACA	TGGCGATAGC	TAGACTGGGC	GGTTTTATGG	3900
25	ACAGCAAGCG	AACCGGAATT	GCCAGCTGGG	GCGCCCTCTG	GTAAGGTTGG	GAAGCCCTGC	3960
25	AAAGTAAACT	GGATGGCTTT	CTTGCCGCCA	AGGATCTGAT	GGCGCAGGGG	ATCAAGATCT	4020
	GATCAAGAGA	CAGGATGAGG	ATCGTTTCGC	ATGATTGAAC	AAGATGGATT	GCACGCAGGT	4080
30	TCTCCGGCCG	CTTGGGTGGA	GAGGCTATTC	GGCTATGACT	GGGCACAACA	GACAATCGGC	4140
	TGCTCTGATG	CCGCCGTGTT	CCGGCTGTCA	GCGCAGGGGC	GCCCGGTTCT	TTTTGTCAAG	4200
25	ACCGACCTGT	CCGGTGCCCT	GAATGAACTG	CAGGACGAGG	CAGCGCGGCT	ATCGTGGCTG	4260
35				CTCGACGTTG GATCTCCTGT			4320 4380
40	GAGAAAGTAT	CCATCATGGC	TGATGCAATG	CGGCGGCTGC	ATACGCTTGA	TCCGGCTACC	4440
40	TGCCCATTCG	ACCACCAAGC	GAAACATCGC	ATCGAGCGAG	CACGTACTCG	GATGGAAGCC	4500
	GGTCTTGTCG	ATCAGGATGA	TCTGGACGAA	GAGCATCAGG	GGCTCGCGCC	AGCCGAACTG	4560
45	TTCGCCAGGC	TCAAGGCGCG	CATGCCCGAC	GGCGAGGATC	TCGTCGTGAC	CCATGGCGAT	4620
	GCCTGCTTGC	CGAATATCAT	GGTGGAAAAT	GGCCGCTTTT	CTGGATTCAT	CGACTGTGGC	4680
50	CGGCTGGGTG	TGGCGGACCG	CTATCAGGAC	ATAGCGTTGG	CTACCCGTGA	TATTGCTGAA	4740
30	GAGCTTGGCG	GCGAATGGGC	TGACCGCTTC	CTCGTGCTTT	ACGGTATCGC	CGCTCCCGAT	4800
	TCGCAGCGCA	TCGCCTTCTA	TCGCCTTCTT	GACGAGTTCT	TCTGAGCGGG	ACTCTGGGGT	4860
55	TCGAAATGAC	CGACCAAGCG	ACGCCCAACC	TGCCATCACG	AGATTTCGAT	TCCACCGCCG	4920
	CCTTCTATGA	AAGGTTGGGC	TTCGGAATCG	TTTTCCGGGA	CGGAATTCGT	AATCTGCTGC	4980
60	TTGCAAACAA	AAAAACCACC	GCTACCAGCG	GTGGTTTGTT	TGCCGGATCA	AGAGCTACCA	5040
υυ	ACTCTTTTTC	CGAAGGTAAC	TGGCTTCAGC	AGAGCGCAGA	TACCAAATAC	TGTCCTTCTA	5100
	GTGTAGCCGT	AGTTAGGCCA	CCACTTCAAG	AACTCTGTAG	CACCGCCTAC	ATACCTCGCT	5160
65	CTGCTAATCC	TGTTACCAGT	GGCTGCTGCC	AGTGGCGATA	AGTCGTGTCT	TACCGGGTTG	5220

	GACTCAAGAC	GATAGTTACC	GGATAAGGCG	CAGCGGTCGG	GCTGAACGGG	GGGTTCGTGC	5280
	ACACAGCCCA	GCTTGGAGCG	AACGACCTAC	ACCGAACTGA	GATACCTACA	GCGTGAGCAT	5340
5	TGAGAAAGCG	CCACGCTTCC	CGAAGGGAGA	AAGGCGGACA	GGTATCCGGT	AAGCGGCAGG	5400
	GTCGGAACAG	GAGAGCGCAC	GAGGGAGCTT	CCAGGGGGAA	ACGCCTGGTA	TCTTTATAGT	5460
10	CCTGTCGGGT	TTCGCCACCT	CTGACTTGAG	CGTCGATTTT	TGTGATGCTC	GTCAGGGGG	5520
10	CGGAGCCTAT	GGAAAAACGC	CAGCAACGCC	GAGATGCGCC	GCCTCGAGTA	CACCTGCGTC	5580
	ATGCTGAGAC	CCTCAAGCCT	CACTAAAAGG	GTCCCTGCCT	AGTTCTGTTT	ACTAATCTGC	5640
15	CTTATTCTGT	TTTTGTTCCC	ATGTTAAAGA	TAGAGTAAAT	GCAGTATTCT	CCACATAGAG	5700
	ATATAGACTT	CTGAAATTCT	AAGATTAGAA	TTATTTACAA	GAAGAAGTGG	GGAA	5754
20	(2) INFORMA	ATION FOR SE	EQ ID NO:20	:			
20	 (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 4958 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single 						
25		(D) TOPOLOG					

(ii) MOLECULE TYPE: DNA (genomic)

30

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:20:

35	AGGCGGGGAC	AGCAGTGCAG	TGGTGGACAG	AAAGCAAGTG	ATCTAGGCCA	GCAGCCTCCC	60
33	TAAAGGGACT	TCAGCCCACA	AAGCCAAACT	TGTGGCTTTA	ATACAAGCTC	TGTAAATGGT	120
	ААААААААА	AAGTCTACAC	GGACAGCAGG	TATGCTCTTG	CCACTGTACA	GAGCAATATA	180
40	CAGACAAAGA	GAACTGTTGA	CATCTGCAGA	GAAAGACCTA	AGATGCTGTG	GCTAAAAGAA	240
	ATCAGATGGC	AAATCTAACC	GCCCAGGCAT	CCTAAAGAGC	AATGATCCTG	ACAGTCTGAA	300
45	GACTATCAAG	TTATAGACAA	ATTAAGACTG	GTAAAAAAAA	CCCTGTATAA	AATAGTAAAA	360
43	ACTGAAAAAA	GAAAACTAGT	CCTCTCATGA	GAAGACAGAC	CTGACATCTA	CTGAAAAATA	420
	GACTTTACTG	GAAAAAATAT	GTGTATGAAT	ACCTTCTAGT	TTTTGTGAAC	GTTCTCAAGA	480
50	TGGATAAAAG	CTTTTCCTTG	TAAAACGAGA	CTGATCAGAT	AGTCATCAAG	AAGATTGTTA	540
	AAGAAAATTT	TCCAAGGTTC	GGAGTGCCAA	AAGCAATAGT	GTCAGATAAT	GGTCCTGCCT	600
55	TTGTTGCCCA	GGTAAGTCAG	GGTGTGGCCA	AGTATTTAGA	GGTCAAATGA	AAATTCCATT	660
33	GTGTGTACAG	ACCTCAGAGC	TCAGGAAAGA	TAAAAAAGAA	TAAATAAAAC	TCTAAACAGA	720
	CCTTGACAAA	ATTAATCCTA	GAGACTGGCA	CAGACTTACT	TGGTACTCCT	TCCCCTTGCC	780
60	CTATTTAGAA	CTGAGAATAC	TCCCTCTTGA	TTCGGTTTTA	CTCTTTTTAA	GATCCTTTAT	840
	GGGGCTCCTA	TGCCATCACT	GTCTTAAATG	ATGTGTTTAA	ACCTATGTTG	TTATAATAAT	900
65	GATCTATATG	TTAAGTTAAA	AGGCTTGCAG	GTGGTGCAGA	AAGAAGTCTG	GTCACAACTG	960
05	GCTACAGTGA	ACAAGCTGGG	TACCCCAAGG	ACATCTTACC	AGTTCCAGCC	AGAGATCTGA	1020

	TCTACGATCC	CCGGGTCGAC	CCGGGTCGAC	CCTGTGGAAT	GTGTGTCAGT	TAGGGTGTGG	1080
	AAAGTCCCCA	GGCTCCCCAG	CAGGCAGAAG	TATGCAAAGC	ATGCATCTCA	ATTAGTCAGC	1140
5	AACCAGGTGT	GGAAAGTCCC	CAGGCTCCCC	AGCAGGCAGA	AGTATGCAAA	GCATGCATCT	1200
	CAATTAGTCA	GCAACCATAG	TCCCGCCCCT	AACTCCGCCC	ATCCCGCCCC	TAACTCCGCC	1260
10	CAGTTCCGCC	CATTCTCCGC	CCCATGGCTG	ACTAATTTTT	TTTATTTATG	CAGAGGCCGA	1320
	GGCCGCCTCG	GCCTCTGAGC	TATTCCAGAA	GTAGTGAGGA	GGCTTTTTTG	GAGGCCTAGG	1380
1.5	CTTTTGCAAA	AAGCTTCACG	CAGGCAGAAG TATGCAAAGC ATGCATCTCA ATTAGTCAGC CAGGCTCCCC AGCAGGCAGA AGTATGCAAA GCATGCATCT TCCCGCCCCT AACTCCGCCC ATCCCGCCCC TAACTCCGCC CCCATGGCTG ACTAATTTTT TTTATTTATG CAGAGGCCGA TATTCCAGAA GTAGTGAGGA GGCTTTTTTG GAGGCCTAGG TATTCCAGAA GTAGTGAGGA GGCTTTTTTG GAGGCCTAGG TATTCCAGAA CACTCAGGGC GCAAGGGCTG CTAAAGGAAG CTGCCGCAAGA CACTCAGGGC GCAAGGGCTG CTAAAGGAAG CTGCCGCAGAAA CGGTGCTGAC CCCGGATGAA TGTCAGCTAC AAACGCAAGC GCAAAGAGAA AGCAGGTAGC TTGCAGTGGG GGTTGGGAAG CCCTGCAAAG TAAACTGGAT GGCTTTCTTG GGATGGGAAG CCCTGCAAAG TAAACTGGAT GGCTTTCTTG GGATTGCAC GCAGGTTCT CGGCCGCTTG GGTGGAGAGG ACAACAGACA ATCGGCTGCT CTGATGCCGC CGTGTTCCGG ACAACAGACA ATCGGCTGCT CTGATGCCGC CGTGTTCCGG GGTTCTTTTT GTCAAGACCG ACCTGTCCGG TGCCCTGAAT GCGGCTATCG TGGCTGGCCA CGACGGGCT TCCTTGCGCA TGAAGCGGGA AGGGACTGGC TGCTATTGGG CGAAGTGCCG CTGACCTTGCT CCTGCCGAGA AAGTATCCAT CATGGCTGAT GCTTGATCCG GCTACCTGC CATTCGACCA CCAAGCGAAA CACCAGACG AAGGGACTGC TGCTATTGGG CGAAGTGCCG CCTGATCCG GCAACTGCC CATTCGACCA CCAAGCGAAA CACCAGACG AAGGACTGCC CATTCGACCA CCAAGCGAAA CCCTGGATCCG GAACTGTCC CATTCGACCA CCAAGCGAAA CCCGCCCACC GAACTGTCC CATTCGACCA CCAAGCGAAA CCGCCCACCC GAACTGTCC CATTCGCCAA GGCACGCCATG CCGCCCACCC GAACTGTCC CCTTGCCGAA TATCATTGGT CCGCGCCACC GAACTGTCC CATTCGCCAA TATCATTGGT CCGGCCCACC GAACTGTCC CATTCGCCAA TATCATTGGT CCGGGCCACC GAACTGTCC CAGGCTCAA GGCGCCCTAT CCGGGCCACC CGAACTGTCC CCTTGCCGAA TATCATTGGT CCGTGATATT GCTGAAGAGC TTGGCGGCAA ATGGCCTAC CCGTGATATT GCTGAAGAGC TTGGCGGCAA ATGGCCTAC CCGTGATATT GCTGAAGAGC TTGGCGGCAA ATGGCCTAC CCGGGCCCT CCGGTTCGC AAGCACACC CAAGCGACGC TTCCATTCCA CCGCCGCCTT CTATGAAAAAAAAAAAAA				
15	CGGAACACGT	AGAAAGCCAG	TCCGCAGAAA	CGGTGCTGAC	CCCGGATGAA	TGTCAGCTAC	1500
	TGGGCTATCT	GGACAAGGGA	AAACGCAAGC	GCAAAGAGAA	AGCAGGTAGC	TTGCAGTGGG	1560
20	CTTACATGGC	GATAGCTAGA	CTGGGCGGTT	TTATGGACAG	CAAGCGAACC	GGAATTGCCA	1620
	GCTGGGGCGC	CCTCTGGTAA	GGTTGGGAAG	CCCTGCAAAG	TAAACTGGAT	GGCTTTCTTG	1680
25	CCGCCAAGGA	TCTGATGGCG	CAGGGGATCA	AGATCTGATC	AAGAGACAGG	ATGAGGATCG	1740
25	TTTCGCATGA	TTGAACAAGA	TGGATTGCAC	GCAGGTTCTC	CGGCCGCTTG	GGTGGAGAGG	1800
	CTATTCGGCT	ATGACTGGGC	ACAACAGACA	ATCGGCTGCT	CTGATGCCGC	CGTGTTCCGG	1860
30	CTGTCAGCGC	AGGGGCGCCC	GGTTCTTTTT	GTCAAGACCG	ACCTGTCCGG	TGCCCTGAAT	1920
	GAACTGCAGG	ACGAGGCAGC	GCGGCTATCG	TGGCTGGCCA	CGACGGGCGT	TCCTTGCGCA	1980
35	GCTGTGCTCG GGGCAGGATC	ACGTTGTCAC TCCTGTCATC	TGAAGCGGGA TCACCTTGCT	AGGGACTGGC CCTGCCGAGA	TGCTATTGGG AAGTATCCAT	CGAAGTGCCG CATGGCTGAT	
	GCAATGCGGC	GGCTGCATAC	AAAGTCCC CAGGCTCCCC AGCAGCAGA AGTATGCAAA GCATGCATCT 1200 AACCATAG TCCCGCCCCT AACTCCGCCC ATCCCGCCC TAACTCCGCC 1260 PTCTCCGC CCCATGGCTG ACTAATTTT TTTATTTATT CAGAGGCCGA 1320 PTCTCAGC TATTCCAGAA GTAGTGAGGA GGCTTTTTTG GAGGCCTAGG 1380 PTCTCAGC TATTCCAGAA GTAGTGAGGA GGCTTTTTTG GAGGCCTAGG 1380 PTCTCACG CTGCCGCAGA CACTCAGGGC GCAAGGGCTG CTAAAGGAAG AAAAGCCAG TCCGCAGAAA CGGTGCTGAC CCCGGATGAA TGTCAGGTAC ACAAGGGA AAACGCAAGC GCAAAGAGAA AGCAGGTAGC TTGCAGTGGG 1500 ACAAGGGA AAACGCAAGC CCCTGCAAAG TAAACTGGAT GGCTTTCTTG 1680 ACAAGGGA AAACGCAAG CCCTGCAAAG TAAACTGGAT GGCTTTCTTG 1680 AGACAAGA TGGATTGCAC GCAGGTTCTC CGGCCGCTTG GGTGGAGAGG 1800 AGACAGAGA ATCGGCTGCT CTGATGCCGC CGTGTTCCGG 1800 AGACAGAGA ATCGGCTGCT CTGATGCCGC CGTGTTCCGG 1800 AGACAGAGA ACACAGACA ATCGGCTGCT CTGATGCCGC CGTGTTCCGG 1800 AGACAGAGA CGGGCTATCC TGGCTGGCCA CGACGGGCGT TCCTTGCGCA 1920 AGAGCAGC GCGGCTATCC TGGCTGGCCA CGACGGGCGT TCCTTGCGCA 1980 AGTTGTCAC TGAAGCCGGA AGGGACTGGC TGCTATTGGG CGAAGTGCCG CTGCATAC GCTTGATCCG GCTACCTGCC CATTCGACCA CCAAGCGAAA 2100 ACCAGGCACC TACTCGGATG GAAGCCGGTC TTGTCGACA CCAAGCGAAA 2100 ACCAGGGCCC GCTTGATCCG GCTACCTGCC CATTCGACCA CCAAGCGAAA 2100 ACCAGGGCCC GGCCCAGCC GAACTGTCC CATTCGACCA CCAAGCGAAA 2220 ACCAGGGCCT CCGCCCAGCC GAACTGTCC CATTCGACCA CCAAGCGAAA 2240 ACCAGGGCCC GCCCAGCC GAACTGTCC CATTCGACCA CCAAGCGAAA 2250 ACCAGGGCT CCGGCCAGCC GAACTGTCC CTTGCCGAA TATCATGGTG ATTCTTCTGG ATTCATCGAC TGTGGCCGCC TGGGTTGAC CCAAGCGACC ATTCTTCTGG ATTCATCGAC TGGGCGTCCT TGTCGCGAA TATCATGGTG ATTCTTCTGG ATTCATCGAC TGGGGTTCGA AATGACCGAC CAAGCGACGC ATTCTTCTGG TATCGCCGCT CCGATTCGC ACAGCGCACC CAAGCGACGC ATTCTTCTGG TATCGCCGCT CCGATTCGC AGCGCATCGC CTTCTATCGC ACACGAGAT TTCGATTCCA CCGCCCTT CTATGAAAGG TTGGCCTCAC ACCAGAGAT TTCGATTCCA CCGCCCCTT CTATGAAAAAAAA ACCACCGCTA ACCACGAGAT TTCGATTCCA CCGCCCCTT CTATGAAAGG TTGGCCTCCC ACCAGAGATACC AAATACTGC CTTCTAGTG AACCAAACA AACCACAC				
40	CATCGCATCG	AGCGAGCACG	TACTCGGATG	GAAGCCGGTC	TTGTCGATCA	GGATGATCTG	2220
40	GACGAAGAGC	ATCAGGGGCT	CGCGCCAGCC	GAACTGTTCG	CCAGGCTCAA	GGCGCGCATG	2280
	CCCGACGGCG	AGGATCTCGT	CGTGACCCAT	GGCGATGCCT	GCTTGCCGAA	TATCATGGTG	2340
45	GAAAATGGCC	GCTTTTCTGG	ATTCATCGAC	TGTGGCCGGC	TGGGTGTGGC	GGACCGCTAT	2400
	CAGGACATAG	CGTTGGCTAC	CCGTGATATT	GCTGAAGAGC	TTGGCGGCGA	ATGGGCTGAC	2460
50	CGCTTCCTCG	TGCTTTACGG	TATCGCCGCT	CCCGATTCGC	AGCGCATCGC	CTTCTATCGC	2520
50	CTTCTTGACG	AGTTCTTCTG	AGCGGGACTC	TGGGGTTCGA	A AATGACCGAC	CAAGCGACGC	2580
	CCAACCTGCC	ATCACGAGAT	TTCGATTCCA	CCGCCGCCTT	CTATGAAAGG	TTGGGCTTCG	2640
55	GAATCGTTTT	CCGGGACGGA	ATTCGTAATC	TGCTGCTTGC	С АААСААААА	ACCACCGCTA	2700
	CCAGCGGTGG	TTTGTTTGCC	GGATCAAGAG	CTACCAACTO	C TTTTTCCGAA	GGTAACTGGC	2760
60	TTCAGCAGAG	CGCAGATACC	AAATACTGTC	CTTCTAGTG	r AGCCGTAGTT	AGGCCACCAC	2820
θU	TTCAAGAACT	CTGTAGCACC	GCCTACATAC	CTCGCTCTG	C TAATCCTGTT	ACCAGTGGCT	2880
	GCTGCCAGTG	GCGATAAGTC	GTGTCTTACC	GGGTTGGAC	r caagacgata	GTTACCGGAT	2940
65	AAGGCGCAGC	GGTCGGGCTG	AACGGGGGG	TCGTGCACAC	C AGCCCAGCTT	GGAGCGAACG	3000

	ACCTACACCG	AACTGAGATA	CCTACAGCGT	GAGCATTGAG	AAAGCGCCAC	GCTTCCCGAA	3060
	GGGAGAAAGG	CGGACAGGTA	TCCGGTAAGC	GGCAGGGTCG	GAACAGGAGA	GCGCACGAGG	3120
. 5	GAGCTTCCAG	GGGGAAACGC	CTGGTATCTT	TATAGTCCTG	TCGGGTTTCG	CCACCTCTGA	3180
	CTTGAGCGTC	GATTTTTGTG	ATGCTCGTCA	GGGGGGCGGA	GCCTATGGAA	AAACGCCAGC	3240
10	AACGCCGAGA	TGCGCCGCCT	CGAGTACACC	TGCGTCATGC	TGAGACCCTC	AAGCCTCACT	3300
10	AAAAGGGTCC	CTGCCTAGTT	CTGTTTACTA	ATCTGCCTTA	TTCTGTTTTT	GTTCCCATGT	3360
	TAAAGATAGA	GTAAATGCAG	TATTCTCCAC	ATAGAGATAT	AGACTTCTGA	AATTCTAAGA	3420
15	TTAGAATTAT	TTACAAGAAG	AAGTGGGGAA	TGAAGAATAA	AAAATTACTG	GCCTCTTGTG	3480
	AGAACATGAA	CTTTCACCTC	GGAGCCCACC	CCCTCCCATC	TGGAAAACAT	ACTTGAGAAA	3540
20	AACATTTTCT	GGAACAACCA	CAGAATGTTT	CAACAGGCCA	GATGTATTGC	CAAACACAGG	3600
20	ATATGACTCT	TTGGTTGAGT	AAATTTGTGG	TTGTTAAACT	TCCCCTATTC	CCTCCCCATT	3660
	CCCCCTCCCA	GTTTGTGGTT	TTTTCCTTTA	AAAGCTTGTG	AAAAATTTGA	GTCGTCGTCG	3720
25	AGACTCCTCT	ACCCTGTGCA	AAGGTGTATG	AGTTTCGACC	CCAGAGCTCT	GTGTGCTTTC	3780
	TGTTGCTGCT	TTATTTCGAC	CCCAGAGCTC	TGGTCTGTGT	GCTTTCATGT	CGCTGCTTTA	3840
20	TTAAATCTTA	CCTTCTACAT	TTTATGTATG	GTCTCAGTGT	CTTCTTGGGT	ACGCGGCTGT	3900
30	CCCGGGACTT	GAGTGTCTGA	GTGAGGGTCT	TCCCTCGAGG	GTCTTTCATT	TGGTACATGG	3960
	GCCGGGAATT	CGAGAATCTT GCCGGGAAAC	TCATTTGGTG	CATTGGCCGG	GAATTCGAAA	ATCTTTCATT	4020 4080
35		TCTGTTTTGG					4140
						CCGAGAGGGA	4200
40		GATAAGGATA					4260
40		GAACAGGGGA					4320
						TGCGAGATCG	4380
45		TCCCACCTCG					4440
		GGATCGTGGG					4500
50						CGTGCAGAGG	4560
30						TTTTGTCTTA	4620
						TCTGTGTCCA	4680
55						CGTTTGTTTT	4740
						CGGTTTCTGT	4800
60						CTGTTTTTAA	4860
00						CCACTTCCTT	4920
		AGCTGCCCTI					4958
65		AGCIGCCOIL					
	(2) INLONG			•			

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 7080 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

10

5

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:21:

15	GAATACAAGC	TTGCATGCCT	GCAGGTCGAC	TCTAGAGGAT	CTTGAAGAAT	AAAAATTAC	60
	TGGCCTCTTG	TGAGAACATG	AACTTTCACC	TCGGAGCCCA	CCCCTCCCA	TCTGGAAAAC	120
20	ATACTTGAGA	AAAACATTTT	CTGGAACAAC	CACAGAATGT	TTCAACAGGC	CAGATGTATT	180
20	GCCAAACACA	GGATATGACT	CTTTGGTTGA	GTAAATTTGT	GGTTGTTAAA	CTTCCCCTAT	240
	TCCCTCCCCA	TTCCCCCTCC	CAGTTTGTGG	TTTTTTCCTT	TAAAAGCTTG	TGAAAAATTT	300
25	GAGTCGTCGT	CGAGACTCCT	CTACCCTGTG	CAAAGGTGTA	TGAGTTTCGA	CCCCAGAGCT	360
	CTGTGTGCTT	TCTGTTGCTG	CTTTATTTCG	ACCCCAGAGC	TCTGGTCTGT	GTGCTTTCAT	420
	GTCGCTGCTT	TATTAAATCT	TACCTTCTAC	ATTTTATGTA	TGGTCTCAGT	GTCTTCTTGG	480
30	GTACGCGGCT	GTCCCGGGAC	TTGAGTGTCT	GAGTGAGGGT	CTTCCCTCGA	GGGTCTTTCA	540
	TTTGGTACAT	GGGCCGGGAA	TTCGAGAATC	TTTCATTTGG	TGCATTGGCC	GGGAATTCGA	600
35	AAATCTTTCA	TTTGGTGCAT	TGGCCGGGAA	ACAGCGCGAC	CACCCAGAGG	TCCTAGACCC	660
	ACTTAGAGGT	AAGATTCTTT	GTTCTGTTTT	GGTCTGATGT	CTGTGTTCTG	ATGTCTGTGT	720
	TCTGTTTCTA	AGTCTGGTGC	GATCGCAGTT	TCAGTTTTGC	GGACGCTCAG	TGAGACCGCG	780
40	CTCCGAGAGG	GAGTGCGGGG	TGGATAAGGA	TAGACGTGTC	CAGGTGTCCA	CCGTCCGTTC	840
	GCCCTGGGAG	ACGTCCCAGG	AGGAACAGGG	GAGGATCAGG	GACGCCTGGT	GGACCCCTTT	900
45	GAAGGCCAAG	AGACCATTTG	GGGTTGCGAG	ATCGTGGGTT	CGAGTCCCAC	CATCGATGGT	960
	GCAGAGGGTC	TCAATTGGCC	GGCCTTAGAA	TTACGGATCT	AGCATGATTG	AACAAGATGG	1020
	ATTGCACGCA	GGTTCTCCGG	CCGCTTGGGT	GGAGAGGCTA	TTCGGCTATG	ACTGGGCACA	1080
50	ACAGACAATC	GGCTGCTCTG	ATGCCGCCGT	GTTCCGGCTG	TCAGCGCAGG	GGCGCCCGGT	1140
	TCTTTTTGTC	AAGACCGACC	TGTCCGGTGC	CCTGAATGAA	CTGCAGGACG	AGGCAGCGCG	1200
55	GCTATCGTGG	CTGGCCACGA	CGGGCGTTCC	TTGCGCAGCT	GTGCTCGACG	TTGTCACTGA	1260
	AGCGGGAAGG	GACTGGCTGC	TATTGGGCGA	AGTGCCGGGG	CAGGATCTCC	TGTCATCTCA	1320
	CCTTGCTCCT	GCCGAGAAAG	TATCCATCAT	GGCTGATGCA	ATGCGGCGGC	TGCATACGCT	1380
60	TGATCCGGCT	ACCTGCCCAT	TCGACCACCA	AGCGAAACAT	CGCATCGAGC	GAGCACGTAC	1440
	TCGGATGGAA	GCCGGTCTTG	TCGATCAGGA	TGATCTGGAC	GAAGAGCATC	AGGGGCTCGC	1500
65	GCCAGCCGAA	CTGTTCGCCA	GGCTCAAGGC	GCGCATGCCC	GACGGCGAGG	ATCTCGTCGT	1560

	GACCCATGGC	GATGCCTGCT	TGCCGAATAT	CATGGTGGAA	AATGGCCGCT	TTTCTGGATT	1620
	CATCGACTGT	GGCCGGCTGG	GTGTGGCGGA	CCGCTATCAG	GACATAGCGT	TGGCTACCCG	1680
. 5	TGATATTGCT	GAAGAGCTTG	GCGGCGAATG	GGCTGACCGC	TTCCTCGTGC	TTTACGGTAT	1740
	CGCCGCTCCC	GATTCGCAGC	GCATCGCCTT	CTATCGCCTT	CTTGACGAGT	TCTTCTGAGC	1800
	GGGACTCTGG	GGTTCGTAAT	GACCGACCAA	GCGACGCCCA	ACCTGCCATC	ACGAGATTTC	1860
10	GATTCCACCG	CCGCCTTCTA	TGAAAGGTTG	GGCTTCGGAG	TTAGCTTGTT	TCTTTACTGT	1920
	TTGTCAATTC	TATTATTTCA	ATACAGAACA	ATAGCTTCTA	TAACTGAAAT	ATATTTGCTA	1980
15	TTGTATATTA	TGATTGTCCC	TCGAACCATG	AACACTCCTC	CAGCTGAATT	TCACAATTCC	2040
	TCTGTCATCT	GCCAGGCCAT	TAAGTTATTC	ATGGAAGATC	TTTGAGGAAC	ACTGCAAGTT	2100
20	САТАТСАТАА	ACACATTTGA	AATTGAGTAT	TGTTTTGCAT	TGTATGGAGC	TATGTTTTGC	2160
20	TGTATCCTCA	GAAAAAAAGT	TTGTTATAAA	GCATTCACAC	CCATAAAAAG	ATAGATTTAA	2220
	ATATTCCAGC	TATAGGAAAG	AAAGTGCGTC	TGCTCTTCAC	TCTAGTCTCA	GTTGGCTCCT	2280
25	TCACATGCAT	GCTTCTTTAT	TTCTCCTATT	TTGTCAAGAA	AATAATAGGT	CACGTCTTGT	2340
	TCTCACTTAT	GTCCTGCCTA	GCATGGCTCA	GATGCACGTT	GTAGATACAA	GAAGGATCAA	2400
20	ATGAAACAGA	CTTCTGGTCT	GTTACTACAA	CCATAGTAAT	AAGCACACTA	ACTAATAATT	2460
30	GCTAATTATG TATCTGGTTG	TTTTCCATCT TAACTGAAGC	CTAAGGTTCC TCAATGGAAC	CACATTTTTC ATGAGCAATA	TGTTTTCTTA TTTCCCAGTC	AAGATCCCAT TTCTCTCCCA	2520 2580
35	TCCAACAGTC	CTGATGGATT	AGCAGAACAG	GCAGAAAACA	CATTGTTACC	CAGAATTAAA	2640
33	AACTAATATT	TGCTCTCCAT	TCAATCCAAA	ATGGACCTAT	TGAAACTAAA	ATCTAACCCA	2700
	ATCCCATTAA	ATGATTTCTA	TGGCGTCAAA	GGTCAAACTT	CTGAAGGGAA	CCTGTGGGTG	2760
40	GGTCACAATT	CAGGCTATAT	ATTCCCCAGG	GCTCAGCCAG	TGTCTGTACA	TACACAACGG	2820
	ATCCTGTGGA	CAGCTCACCT	AGCTGCAATG	GCTACAGGCT	CCCGGACGTC	CCTGCTCCTG	2880
45	GCTTTTGGCC	TGCTCTGCCT	GCCCTGGCTT	CAAGAGGGCA	GTGCCTTCCC	AACCATTCCC	2940
43	TTATCCAGG	TTTTTGACAA	CGCTATGCTC	CGCGCCCATC	GTCTGCACCA	GCTGGCCTTT	3000
	GACACCTAC	AGGAGTTTGA	AGAAGCCTAT	ATCCCAAAG	AACAGAAGTA	TTCATTCCTG	3060
50	CAGAACCCCC	AGACCTCCCT	CTGTTTCTC	A GAGTCTATTO	CGACACCCTC	CAACAGGGAG	3120
	GAAACACAA	C AGAAATCCAA	CCTAGAGCT	CTCCGCATC	CCCTGCTGCT	CATCCAGTCG	3180
55	TGGCTGGAG	CCGTGCAGTT	CCTCAGGAG	r GTCTTCGCC	A ACAGCCTGGT	GTACGGCGCC	3240
33	TCTGACAGC	A ACGTCTATGA	CCTCCTAAA	G GACCTAGAGO	S AAGGCATCCA	AACGCTGATG	3300
	GGGAGGCTG	G AAGATGGCAG	CCCCCGGAC	r GGGCAGATC	T TCAAGCAGAC	CTACAGCAAG	3360
60						GCTCTACTGC	3420
						CCGCTCTGTG	3480
65						CCAGTGCCTC	3540
CO	TCCTGGCCC	T GGAAGTTGCC	ACTCCAGTG	C CCACCAGCC	T TGTCCTAATA	AAATTAAGTT	3600

	GCATCAAAAA	ааааааааа	CTAGCGGCCG	CTAGACTTCT	GAAATTCTAA	GATTAGAATT	3660
	ATTTACAAGA	AGAAGTGGGG	AATGAAGAAT	AAAAAATTAC	TGGCCTCTTG	TGAGAACATG	3720
5	AACTTTCACC	TCGGAGCCCA	CCCCTCCCA	TCTGGAAAAC	ATACTTGAGA	AAAACATTTT	3780
	CTGGAACAAC	CACAGAATGT	TTCAACAGGC	CAGATGTATT	GCCAAACACA	GGATATGACT	3840
10	CTTTGGTTGA	GTAAATTTGT	GGTTGTTAAA	CTTCCCCTAT	TCCCTCCCCA	TTCCCCCTCC	3900
	CAGTTTGTGG	TTTTTTCCTT	TAAAAGCTTG	TGAAAAATTT	GAGTCGTCGT	CGAGACTCCT	3960
1.5	CTACCCTGTG	CAAAGGTGTA	TGAGTTTCGA	CCCCAGAGCT	CTGTGTGCTT	TCTGTTGCTG	4020
15	CTTTATTTCG	ACCCCAGAGC	TCTGGTCTGT	GTGCTTTCAT	GTCGCTGCTT	TATTAAATCT	4080
	TACCTTCTAC	ATTTTATGTA	TGGTCTCAGT	GTCTTCTTGG	GTACGCGGCT	GTCCCGGGAC	4140
20	TTGAGTGTCT	GAGTGAGGGT	CTTCCCTCGA	GGGTCTTTCA	TTTGGTACAT	GGGCCGGGAA	4200
	TTCGAGAATC	TTTCATTTGG	TGCATTGGCC	GGGAATTCGA	AAATCTTTCA	GATCCCCGGG	4260
25	TACCGAGCTC	GAATTCCGGT	CTCCCTATAG	TGAGTCGTAT	TAATTTCGAT	AAGCCAGCTG	4320
25	CATTAATGAA	TCGGCCAACG	CGCGGGGAGA	GGCGGTTTGC	GTATTGGGCG	CTCTTCCGCT	. 4380
	TCCTCGCTCA	CTGACTCGCT	GCGCTCGGTC	GTTCGGCTGC	GGCGAGCGGT	ATCAGCTCAC	4440
30	TCAAAGGCGG GCAAAAGGCC	TAATACGGTT AGCAAAAGGC	ATCCACAGAA CAGGAACCGT	TCAGGGGATA AAAAAGGCCG	ACGCAGGAAA CGTTGCTGGC	GAACATGTGA GTTTTTCCAT	4500 4560
	AGGCTCCGCC	CCCCTGACGA	GCATCACAAA	AATCGACGCT	CAAGTCAGAG	GTGGCGAAAC	4620
35	CCGACAGGAC	TATAAAGATA	CCAGGCGTTT	CCCCTGGAA	GCTCCCTCGT	GCGCTCTCCT	4680
	GTTCCGACCC	TGCCGCTTAC	CGGATACCTG	TCCGCCTTTC	TCCCTTCGGG	AAGCGTGGCG	4740
40	CTTTCTCATA	GCTCACGCTG	TAGGTATCTC	AGTTCGGTGT	AGGTCGTTCG	CTCCAAGCTG	4800
40	GGCTGTGTGC	ACGAACCCCC	CGTTCAGCCC	GACCGCTGCG	CCTTATCCGG	TAACTATCGT	4860
	CTTGAGTCCA	ACCCGGTAAG	ACACGACTTA	TCGCCACTGG	CAGCAGCCAC	TGGTAACAGG	4920
45	ATTAGCAGAG	CGAGGTATGT	AGGCGGTGCT	ACAGAGTTCT	TGAAGTGGTG	GCCTAACTAC	4980
	GGCTACACTA	GAAGGACAGT	ATTTGGTATC	TGCGCTCTGC	TGAAGCCAGT	TACCTTCGGA	5040
50	AAAAGAGTTG	GTAGCTCTTG	ATCCGGCAAA	CAAACCACCG	CTGGTAGCGG	TGGTTTTTTT	5100
50	GTTTGCAAGC	: AGCAGATTAC	GCGCAGAAAA	AAAGGATCTO	AAGAAGATCC	TTTGATCTTT	5160
	TCTACGGGGT	CTGACGCTCA	GTGGAACGAA	AACTCACGTT	AAGGGATTTT	GGTCATGAGA	5220
55	ттатсааааа	GGATCTTCAC	CTAGATCCTT	TTAAATTAA	A AATGAAGTTT	TAAATCAATC	5280
	TAAAGTATAT	ATGAGTAAAC	TTGGTCTGAC	AGTTACCAAT	r GCTTAATCAG	TGAGGCACCT	5340
60	ATCTCAGCGA	A TCTGTCTATT	TCGTTCATC	ATAGTTGCC1	r GACTCCCCGT	CGTGTAGATA	5400
00	ACTACGATAC	GGGAGGGCTT	ACCATCTGG	CCCAGTGCT	G CAATGATACC	GCGAGACCCA	5460
	CGCTCACCGC	CTCCAGATT	ATCAGCAATA	A AACCAGCCA	G CCGGAAGGGC	CGAGCGCAGA	5520
65	AGTGGTCCTC	CAACTTTATO	CGCCTCCAT	CAGTCTATT	A ATTGTTGCCG	GGAAGCTAGA	5580

GTAAGTAGTT CGCCAGTTAA TAGTTTGCGC AACGTTGTTG CCATTGCTAC AGGCATCGTG 5640 GTGTCACGCT CGTCGTTTGG TATGGCTTCA TTCAGCTCCG GTTCCCAACG ATCAAGGCGA 5700 GTTACATGAT CCCCCATGTT GTGCAAAAAA GCGGTTAGCT CCTTCGGTCC TCCGATCGTT 5760 5820 GTCAGAAGTA AGTTGGCCGC AGTGTTATCA CTCATGGTTA TGGCAGCACT GCATAATTCT CTTACTGTCA TGCCATCCGT AAGATGCTTT TCTGTGACTG GTGAGTACTC AACCAAGTCA 5880 10 TTCTGAGAAT AGTGTATGCG GCGACCGAGT TGCTCTTGCC CGGCGTCAAT ACGGGATAAT 5940 ACCGCGCCAC ATAGCAGAAC TTTAAAAGTG CTCATCATTG GAAAACGTTC TTCGGGGGCGA 6000 AAACTCTCAA GGATCTTACC GCTGTTGAGA TCCAGTTCGA TGTAACCCAC TCGTGCACCC 6060 15 AACTGATCTT CAGCATCTTT TACTTTCACC AGCGTTTCTG GGTGAGCAAA AACAGGAAGG 6120 CAAAATGCCG CAAAAAAGGG AATAAGGGCG ACACGGAAAT GTTGAATACT CATACTCTTC 6180 20 6240 CTTTTTCAAT ATTATTGAAG CATTTATCAG GGTTATTGTC TCATGAGCGG ATACATATTT 6300 GAATGTATTT AGAAAAATAA ACAAATAGGG GTTCCGCGCA CATTTCCCCG AAAAGTGCCA 6360 CCTGACGTCT AAGAAACCAT TATTATCATG ACATTAACCT ATAAAAATAG GCGTATCACG 25 AGGCCCTTTC GTCTCGCGCG TTTCGGTGAT GACGGTGAAA ACCTCTGACA CATGCAGCTC 6420 CCGGAGACGG TCACAGCTTG TCTGTAAGCG GATGCCGGGA GCAGACAAGC CCGTCAGGGC 6480 6540 GCGTCAGCGG GTGTTGGCGG GTGTCGGGGC TGGCTTAACT ATGCGGCATC AGAGCAGATT 30 GTACTGAGAG TGCACCATAT CGACGCTCTC CCTTATGCGA CTCCTGCATT AGGAAGCAGC 6600 6660 CCAGTAGTAG GTTGAGGCCG TTGAGCACCG CCGCCGCAAG GAATGGTGCA AGGAGATGGC 35 GCCCAACAGT CCCCGGCCA CGGGGCCTGC CACCATACCC ACGCCGAAAC AAGCGCTCAT 6720 GAGCCCGAAG TGGCGAGCCC GATCTTCCCC ATCGGTGATG TCGGCGATAT AGGCGCCAGC 6780 AACCGCACCT GTGGCGCCGG TGATGCCGGC CACGATGCGT CCGGCGTAGA GGATCTGGCT 6840 40 AGCGATGACC CTGCTGATTG GTTCGCTGAC CATTTCCGGG GTGCGGAACG GCGTTACCAG 6900 AAACTCAGAA GGTTCGTCCA ACCAAACCGA CTCTGACGGC AGTTTACGAG AGAGATGATA 6960 45

50

55

5

(2) INFORMATION FOR SEQ ID NO:22:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 6795 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

60

65

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:22: 60 AATGAAAGAC CCCACCTGTA GGTTTGGCAA GCTAGCTTAA GTAACGCCAT TTTGCAAGGC

GGGTCTGCTT CAGTAAGCCA GATGCTACAC AATTAGGCTT GTACATATTG TCGTTAGAAC

GCGGCTACAA TTAATACATA ACCTTATGTA TCATACACAT ACGATTTAGG TGACACTATA

7020

7080

							-
	ATGGAAAAAT	ACATAACTGA	GAATAGAGAA	GTTCAGATCA	AGGTCAGGAA	CAGATGGAAC	120
	AGCTGAATAT	GGGCCAAACA	GGATATCTGT	GGTAAGCAGT	TCCTGCCCCG	GCTCAGGGCC	180
5	AAGAACAGAT	GGAACAGCTG	AATATGGGCC	AAACAGGATA	TCTGTGGTAA	GCAGTTCCTG	240
	CCCCGGCTCA	GGGCCAAGAA	CAGATGGTCC	CCAGATGCGG	TCCAGCCCTC	AGCAGTTTCT	300
10	AGAGAACCAT	CAGATGTTTC	CAGGGTGCCC	CAAGGACCTG	AAATGACCCT	GTGCCTTATT	360
	TGAACTAACC	AATCAGTTCG	CTTCTCGCTT	CTGTTCGCGC	GCTTCTGCTC	CCCGAGCTCA	420
15	ATAAAAGAGC	CCACAACCCC	TCACTCGGGG	CGCCAGTCCT	CCGATTGACT	GAGTCGCCCG	480
13	GGTACCCGTG	TATCCAATAA	ACCCTCTTGC	AGTTGCATCC	GACTTGTGGT	CTCGCTGTTC	540
	CTTGGGAGGG	TCTCCTCTGA	GTGATTGACT	ACCCGTCAGC	GGGGGTCTTT	CATTTGGGGG	600
20	CTCGTCCGGG	ATCGGGAGAC	CCCTGCCCAG	GGACCACCGA	CCCACCACCG	GGAGGTAAGC	660
	TGGCCAGCAA	CTTATCTGTG	TCTGTCCGAT	TGTCTAGTGT	CTATGACTGA	TTTTATGCGC	720
25	CTGCGTCGGT	ACTAGTTAGC	TAACTAGCTC	TGTATCTGGC	GGACCCGTGG	TGGAACTGAC	780
23	GAGTTCGGAA	CACCCGGCCG	CAACCCTGGG	AGACGTCCCA	GGAGGAACAG	GGGAGGATCA	840
30	GGGACGCCTG TTCGAGTCCC	GTGGACCCCT ACCTCGTGCC	TTGAAGGCCA CAGTTGCGAG	AGAGACCATT ATCGTGGGTT	TGGGGTTGCG CGAGTCCCAC	AGATCGTGGG CTCGTGTTTT	900 960
30	GTTGCGAGAT	CGTGGGTTCG	AGTCCCACCT	CGCGTCTGGT	CACGGGATCG	TGGGTTCGAG	1020
	TCCCACCTCG	TGTTTTGTTG	CGAGATCGTG	GGTTCGAGTC	CCACCTCGCG	TCTGGTCACG	1080
35	GGATCGTGGG	TTCGAGTCCC	ACCTCGTGCA	GAGGGTCTCA	ATTGGCCGGC	CTTAGAGAGG	1140
	CCATCTGATT	CTTCTGGTTT	CTCTTTTTGT	CTTAGTCTCG	TGTCCGCTCT	TGTTGTGACT	1200
40	ACTGTTTTTC	TAAAAATGGG	ACAATCTGTG	TCCACTCCCC	TTTCTCTGAC	TCTGGTTCTG	1260
40	TCGCTTGGTA	ATTTTGTTTG	TTTACGTTTG	TTTTTGTGAG	TCGTCTATGT	TGTCTGTTAC	1320
	TATCTTGTTT	TTGTTTGTGG	TTTACGGTTT	CTGTGTGTGT	CTTGTGTGTC	TCTTTGTGTT	1380
45	CAGACTTGGA	CTGATGACTG	ACGACTGTTT	TTAAGTTATG	CCTTCTAAAA	TAAGCCTAAA	1440
	AATCCTGTCA	GATCCCTATG	CTGACCACTT	CCTTTCAGAT	CAACAGCTGC	CCTTACTCGA	1500
50	GCTCAAGCTT	CGAATTCTGC	AGTCGACGGT	ACCGCGGCCG	СТААСТААТА	GCCCATTCTC	1560
30	CAAGGTACGT	AGCGGGGATC	AATTCCGCCC	CCCCCTAAC	GTTACTGGCC	GAAGCCGCTT	1620
	GGAATAAGGC	CGGTGTGCGT	TTGTCTATAT	GTTATTTTCC	ACCATATTGC	CGTCTTTTGG	1680
55	CAATGTGAGG	GCCCGGAAAC	CTGGCCCTGT	CTTCTTGACG	AGCATTCCTA	GGGGTCTTTC	1740
	CCCTCTCGCC	AAAGGAATGC	AAGGTCTGTT	GAATGTCGTG	AAGGAAGCAG	TTCCTCTGGA	1800
60	AGCTTCTTGA	AGACAAACAA	CGTCTGTAGC	GACCCTTTGC	AGGCAGCGGA	ACCCCCCACC	1860
50	TGGCGACAGG	TGCCTCTGCG	GCCAAAAGCC	ACGTGTATAA	GATACACCTG	CAAAGGCGGC	1920
	ACAACCCCAG	TGCCACGTTG	TGAGTTGGAT	AGTTGTGGAA	AGAGTCAAAT	GGCTCTCCTC	1980
65	AAGCGTÄTTC	AACAAGGGGC	TGAAGGATGC	CCAGAAGGTA	CCCCATTGTA	TGGGATCTGA	2040

	TCTGGGGCCT	CGGTGCACAT	GCTTTACATG	TGTTTAGTCG	AGGTTAAAAA	AACGTCTAGG	2100
	CCCCCGAAC	CACGGGGACG	TGGTTTTCCT	TTGAAAAACA	CGATACGGGA	TCCACCGGTC	2160
5	GCCACCATGG	GTAAAGGAGA	AGAACTTTTC	ACAGGAGTTG	TCCCAATTCT	TGTTGAATTA	2220
	GATGGTGATG	TTAATGGGCA	CAAATTTTCT	GTCAGTGGAG	AGGGTGAAGG	TGATGCAACA	2280
10	TACGGAAAAC	TTACCCTTAA	ATTTATTTGC	ACTACTGGAA	AACTACCTGT	TCCATGGCCA	2340
10	ACACTTGTCA	CTACTTTCAC	TTATGGTGTT	CAATGCTTTT	CAAGATACCC	AGATCATATG	2400
	AAACGGCATG	ACTTTTTCAA	GAGTGCCATG	CCCGAAGGTT	ATGTACAGGA	AAGAACTATA	2460
15	TTTTTCAAAG	ATGACGGGAA	CTACAAGACA	CGTGCTGAAG	TCAAGTTTGA	AGGTGATACC	2520
	CTTGTTAATA	GAATCGAGTT	AAAAGGTATT	GATTTTAAAG	AAGATGGAAA	CATTCTTGGA	2580
20	CACAAATTGG	AATACAACTA	TAACTCACAC	AATGTATACA	TCATGGCAGA	CAAACAAAAG	2640
20	AATGGAACCA	AAGTTAACTT	CAAAATTAGA	CACAACATTG	AAGATGGAAG	CGTTCAACTA	2700
	GCAGACCATT	ATCAACAAAA	TACTCCAATT	GGCGATGGCC	CTGTCCTTTT	ACCAGACAAC	2760
25	CATTACCTGT	CCACACAATC	TGCCCTTTCG	AAAGATCCCA	ACGAAAAGAG	AGACCACATG	2820
	GTCCTTCTTG TCCGGATCTA	AGTTTGTAAC GATAACTGTA	AGCTGCTGGG TCGATGGATC	ATTACACATG CGAAGGCGGG	GCATGGATGA GACAGCAGTG	ACTATACAAG CAGTGGTGGA	2880 2940
30	CAGAAAGCAA	GTGATCTAGG	CCAGCAGCCT	CCCTAAAGGG	ACTTCAGCCC	ACAAAGCCAA	3000
	ACTTGTGGCT	TTAATACAAG	CTCTGTAAAT	GGTAAAAAAA	AAAAAGTCTA	CACGGACAGC	3060
25	AGGTATGCTC	TTGCCACTGT	ACAGAGCAAT	ATACAGACAA	AGAGAACTGT	TGACATCTGC	3120
35	AGAGAAAGAC	CTAAGATGCT	GTGGCTAAAA	GAAATCAGAT	GGCAAATCTA	ACCGCCCAGG	3180
	CATCCTAAAG	AGCAATGATC	CTGACAGTCT	GAAGACTATC	AAGTTATAGA	CAAATTAAGA	3240
40	CTGGTAAAAA	AAACCCTGTA	TAAAATAGTA	AAAACTGAAA	AAAGAAAACT	AGTCCTCTCA	3300
	TGAGAAGACA	GACCTGACAT	CTACTGAAAA	ATAGACTTTA	CTGGAAAAAA	TATGTGTATG	3360
15	AATACCTTCT	AGTTTTTGTG	AACGTTCTCA	AGATGGATAA	AAGCTTTTCC	TTGTAAAACG	3420
45	AGACTGATCA	GATAGTCATC	AAGAAGATTG	TTAAAGAAAA	TTTTCCAAGG	TTCGGAGTGC	3480
	CAAAAGCAAT	AGTGTCAGAT	AATGGTCCTG	CCTTTGTTGC	CCAGGTAAGT	CAGGGTGTGG	3540
50	CCAAGTATTT	AGAGGTCAAA	TGAAAATTCC	ATTGTGTGT.	A CAGACCTCAG	AGCTCAGGAA	3600
	AGATAAAAAA	GAATAAATAA	AACTCTAAAC	AGACCTTGAC	AAAATTAATC	CTAGAGACTG	3660
5.5	GCACAGACTT	ACTTGGTACT	CCTTCCCCTI	GCCCTATTT	A GAACTGAGAA	TACTCCCTCT	3720
55	TGATTCGGTT	TTACTCTTTT	TAAGATCCTI	TATGGGGCT	CTATGCCATC	ACTGTCTTAA	3780
	ATGATGTGTT	TAAACCTATG	TTGTTATAAT	AATGATCTAT	r atgttaagtt	AAAAGGCTTG	3840
60	CAGGTGGTGC	AGAAAGAAGT	CTGGTCACA	CTGGCTACA	G TGAACAAGCI	GGGTACCCCA	3900
	AGGACATCTT	ACCAGTTCC	GCCAGAGATO	TGATCTACGA	A TCCCCGGGTC	GACCCGGGTC	3960
65	GACCCTGTGG	AATGTGTGT	AGTTAGGGT	TGGAAAGTC	C CCAGGCTCCC	CAGCAGGCAG	4020
65	AAGTATGCAA	AGCATGCAT	TCAATTAGTO	AGCAACCAG	G TGTGGAAAGI	CCCCAGGCTC	4080

	CCCAGCAGGC	AGAAGTATGC	AAAGCATGCA	TCTCAATTAG	TCAGCAACCA	TAGTCCCGCC	4140
. 5	CCTAACTCCG	CCCATCCCGC	CCCTAACTCC	GCCCAGTTCC	GCCCATTCTC	CGCCCCATGG	CCCCATGG 4200 CTATTCCA 4260 CTATTCCA 4260 CCTATCCA 4260 CCCCATGC 4320 CCCCATGC 4320 CCCCATGC 4380 AAAACGCA 4440 ACTGGGCG 4500 AGGTTGGG 4560 CCCAGGGGA 4620 ATGGATTG 4680 CCCAACACG 4740 CCGCGCTA 4860 CCCCTT 4980 CCCCTTGAT 5040 CCCCTTGAT 5040 CCCCTTGAT 5040 CCCCTGAT 5220 CCCCGTGAT 5340 CCCCGTGAT 5340 CCCCGTGAT 5520 CCCGTGAT 5520 CCCCTACA 5760 CCGCCTACA 5760 CCGCCTACA 5760 CCGCCTACA 5940 AATCCGGTA 6000
J	CTGACTAATT	TTTTTTTTT	ATGCAGAGGC	CGAGGCCGCC	TCGGCCTCTG	AGCTATTCCA	4260
	GAAGTAGTGA	GGAGGCTTTT	TTGGAGGCCT	AGGCTTTTGC	AAAAAGCTTC	ACGCTGCCGC	4320
10	AAGCACTCAG	GGCGCAAGGG	CTGCTAAAGG	AAGCGGAACA	CGTAGAAAGC	CAGTCCGCAG	4380
	AAACGGTGCT	GACCCCGGAT	GAATGTCAGC	TACTGGGCTA	TCTGGACAAG	GGAAAACGCA	4440
15	AGCGCAAAGA	GAAAGCAGGT	AGCTTGCAGT	GGGCTTACAT	GGCGATAGCT	AGACTGGGCG	4500
13	GTTTTATGGA	CAGCAAGCGA	PATECCEGE CCCTARCTCC GCCCAGTTCC GCCCATTCT CGCCCCATGG 4200 TITTATTT ATGCAGAGGC CGAGGCCGC TCGGCCTCT AGCTATTCCA 4260 AGCCTTT TTGGAGGCCT AGGCTTTTGC AAAAAGCTTC ACGCTGCGC 4320 AGCCAGGG CTGCTAAAAG AAGCGGAACA CGTAGAAAAG CAGTCCGCAG 4380 ACCCCGGAT GAATGTCAGC TACTGGGCTA TCTGGACAAG GGAAAACGCA 4440 AAGCAGGT AGCTTGCAGT GGGCTTACAT GGCGATAGCT AGACTGGGCG 4500 ACAAGCAG ACCGGAATTG CCAGCTGGG CGCCCTCTGG TAAGGTTGGG 4560 ACAAGCAG ACCGGAATTG CCAGCTGGG CGCCCTCTGG TAAGGTTGGG 4620 ACAAGCAG ACGGATATC TTGCCGCCAA GGATCTGATG GCGCAGGGGA 4620 ACAAGAGA AGGATGAGGA TCGTTTCGCA TGATTGAACA AGATGGATTG 4680 ACCGCCCC TTGGGTGGA AGGCTATCG GCCAGAGGGA AGCCCGCTT 4800 ACCTGCCCC TTGGGTGGA AGGCTTACC CGCAGGGGC CCCGGTTCTT 4800 ACCACGCCGC TTGGGTGCAC AAAACATCC AGGACGAGGA AGCCCGGCTA 4800 ACCTGCTATT GGGCGAAGTG CCGGGGCAG ATCTCCTGT CACTGAACCG 4920 ACCTGCTATT GGGCGAAGTG CCGGGGCAG ATCTCCTGT AAAACTAC CATCAACGC GAACCTGTGC TCGACGTTGT AAAACTAC CATCAACGC GAACCTGGAG AACAATCGCA TCGACGTGC ACCCCAACAG AAAACATCC ATCGCTTGAT 5040 ACCATTCGA CCACCAAGCG AAACATCGCA TCGAGCGAGC ACCGCGCCA 5160 ACCAGGCG CTCCTTCAT CCCCTTCC TCGACGTTCC TCGTCGACC 5220 ACCTGCTAT CGGGGACGC ATCCCGACG GCGAGGATC CGTCGGCCA 5160 ACCAGGCC CAAAGGCGC ATCCCCGACG GCGAGGATC CGTCGGCCA 5160 ACCAGGCC CAAAGGGC CACCCCAACC TCGGGTTCT TCGAGCGGG AAAACATCGC TCGGGTTCT TCGGACGTGC 5220 ACCTGCTTCC GAAATACAT GCCCTTCC TCGTGTTCT TCGATCACC 5220 ACCTGCGTTCTTAT CGCCTTCT TCGGACGAC ACCTCTCT TCGAGCGGA 5400 AAAATGACC GAAAGCGC CACCCCAACCT CCGCCTTTC TCGAGCGGA 5400 AAAATGACC GACAAAGCA CCCCCAACCT CCCCTTCTT CGGATTCGC 5400 AAAATGACC GACAAAGCA CCCCCAACCT CCCCTTCTT CGGGATCAC 5500 AAAACAAA AAAACCACCG CTACCAGCG TTTCCAGGACA ACCAAATACT 5500 AAAACAAA AAAACCACCG CTACCAGCG TGCTCTCT TCCGGGAT ACCAAATACT 5700 AAAACAAA AAAACCACCG CTACCAAGCG TGGCGAAAA ACCAAATACT 5700 ATAGCCGTA GTTAGGCCA CACTTCAAGA ACCTGACAA ACCAAATACT 5700 ATAGCCGTA GTTAGGCCA CACTTCAAGA ACCAACTGA GACCGCTACA 5760 ACCAAACAAA AAAACCACG CTACCAAGCG TGGCGAAAA ACCAAATACT 5700 ATAGCCGTA GTTAGGCCA CACTTCAAGA ACCAACTACA 5760 ACCAACAAG ATAGTTACCG GATAAGGCG ACCGCTACA 5760 ACCAACAAGA AAAACCTACG GATAAGGCG ACCGCTAC				
	AAGCCCTGCA	AAGTAAACTG	GATGGCTTTC	TTGCCGCCAA	GGATCTGATG	GCGCAGGGGA	4620
20	TCAAGATCTG	ATCAAGAGAC	AGGATGAGGA	TCGTTTCGCA	TGATTGAACA	AGATGGATTG	4680
	CACGCAGGTT	CTCCGGCCGC	TTGGGTGGAG	AGGCTATTCG	GCTATGACTG	GGCACAACAG	4740
25	ACAATCGGCT	GCTCTGATGC	CGCCGTGTTC	CGGCTGTCAG	CGCAGGGGCG	CCCGGTTCTT	4800
23							
30	GGAAGGGACT	GGCTGCTATT	GGGCGAAGTG	CCGGGGCAGG	ATCTCCTGTC	ATCTCACCTT	4980
30	GCTCCTGCCG	AGAAAGTATC	CATCATGGCT	GATGCAATGC	GGCGGCTGCA	TACGCTTGAT	5040
	CCGGCTACCT	GCCCATTCGA	CCACCAAGCG	AAACATCGCA	TCGAGCGAGC	ACGTACTCGG	5100
35	ATGGAAGCCG	GTCTTGTCGA	TCAGGATGAT	CTGGACGAAG	AGCATCAGGG	GCTCGCGCCA	5160
	GCCGAACTGT	TCGCCAGGCT	CAAGGCGCGC	ATGCCCGACG	GCGAGGATCT	CGTCGTGACC	5220
40	CATGGCGATG	CCTGCTTGCC	GAATATCATG	GTGGAAAATG	GCCGCTTTTC	TGGATTCATC	5280
40	GACTGTGGCC	GGCTGGGTGT	GGCGGACCGC	TATCÀGGACA	TAGCGTTGGC	TACCCGTGAT	5340
	ATTGCTGAAG	AGCTTGGCGG	CGAATGGGCT	GACCGCTTCC	TCGTGCTTTA	CGGTATCGCC	5400
45	GCTCCCGATT	CGCAGCGCAT	CGCCTTCTAT	CGCCTTCTTG	ACGAGTTCTT	CTGAGCGGGA	5460
	CTCTGGGGTT	CGAAATGACC	GACCAAGCGA	CGCCCAACCT	GCCATCACGA	GATTTCGATT	5520
50	CCACCGCCGC	CTTCTATGAA	AGGTTGGGCT	TCGGAATCGT	TTTCCGGGAC	GGAATTCGTA	5580
50	ATCTGCTGCT	TGCAAACAAA	AAAACCACCG	CTACCAGCGG	TGGTTTGTTT	GCCGGATCAA	5,640
	GAGCTACCAA	CTCTTTTTCC	GAAGGTAACT	GGCTTCAGCA	GAGCGCAGAT	ACCAAATACT	5700
55	GTCCTTCTAG	TGTAGCCGTA	GTTAGGCCAC	CACTTCAAGA	ACTCTGTAGC	ACCGCCTACA	5760
	TACCTCGCTC	TGCTAATCCT	GTTACCAGTG	GCTGCTGCCA	GTGGCGATAA	GTCGTGTCTT	5820
60	ACCGGGTTGG	ACTCAAGACG	ATAGTTACCG	GATAAGGCGC	AGCGGTCGGG	CTGAACGGGG	5880
	GGTTCGTGCA	CACAGCCCAG	CTTGGAGCGA	ACGACCTACA	CCGAACTGAG	ATACCTACAG	5940
	CGTGAGCATT	GAGAAAGCGC	CACGCTTCCC	GAAGGGAGAA	AGGCGGACAG	GTATCCGGTA	6000
65	AGCGGCAGGG	TCGGAACAGG	AGAGCGCACG	AGGGAGCTTC	CAGGGGGAAA	CGCCTGGTAT	6060

	CTTTATAGTC CTGTCGGGTT TCGCCACCTC TGACTTGAGC GTCGATTTTT GTGATGCTCG	6120
	TCAGGGGGGC GGAGCCTATG GAAAAACGCC AGCAACGCCG AGATGCGCCG CCTCGAGAAC	6180
5	CCTGGCCCTA TTATTGGGTG GACTAACCAT GGGGGGAATT GCCGCTGGAA TAGGAACAGG	6240
	GACTACTGCT CTAATGGCCA CTCAGCAATT CCAGCAGCTC CAAGCCGCAG TACAGGATGA	6300
10	TCTCAGGGAG GTTGAAAAAT CAATCTCTAA CCTAGAAAAG TCTCTCACTT CCCTGTCTGA	6360
10	AGTTGTCCTA CAGAATCGAA GGGGCCTAGA CTTGTTATTT CTAAAAGAAG GAGGGCTGTG	6420
	TGCTGCTCTA AAAGAAGAAT GTTGCTTCTA TGCGGACCAC ACAGGACTAG TGAGAGACAG	6480
15	CATGGCCAAA TTGAGAGAGA GGCTTAATCA GAGACAGAAA CTGTTTGAGT CAACTCAAGG	6540
	ATGGTTTGAG GGACTGTTTA ACAGATCCCC TTGGTTTACC ACCTTGATAT CTACCATTAT	6600
20	GGGACCCCTC ATTGTACTCC TAATGATTTT GCTCTTCGGA CCCTGCATTC TTAATCGATT	6660
20	AGTCCAATTT GTTAAAGACA GGATATCAGT GGTCCAGGCT CTAGTTTTGA CTCAACAATA	6720
	TCACCAGCTG AAGCCTATAG AGTACGAGCC ATAGATAAAA TAAAAGATTT TATTTAGTCT	6780
25	CCAGAAAAAG GGGGG (2) INFORMATION FOR SEQ ID NO:23:	6795
	(i) SEQUENCE CHARACTERISTICS:	
30	(A) LENGTH: 9093 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear	
35	(ii) MOLECULE TYPE: DNA (genomic)	
40	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:23:	
40	AATGAAAGAC CCCACCTGTA GGTTTGGCAA GCTAGCTTAA GTAACGCCAT TTTGCAAGGC	60
	ATGGAAAAAT ACATAACTGA GAATAGAGAA GTTCAGATCA AGGTCAGGAA CAGATGGAAC	120
45	AGCTGAATAT GGGCCAAACA GGATATCTGT GGTAAGCAGT TCCTGCCCCG GCTCAGGGCC	180
	AAGAACAGAT GGAACAGCTG AATATGGGCC AAACAGGATA TCTGTGGTAA GCAGTTCCTG	240
50	CCCCGGCTCA GGGCCAAGAA CAGATGGTCC CCAGATGCGG TCCAGCCCTC AGCAGTTTCT	300
50	AGAGAACCAT CAGATGTTTC CAGGGTGCCC CAAGGACCTG AAATGACCCT GTGCCTTATT	360
	TGAACTAACC AATCAGTTCG CTTCTCGCTT CTGTTCGCGC GCTTCTGCTC CCCGAGCTCA	420
55	ATAAAAGAGC CCACAACCCC TCACTCGGGG CGCCAGTCCT CCGATTGACT GAGTCGCCCG	480
	GGTACCCGTG TATCCAATAA ACCCTCTTGC AGTTGCATCC GACTTGTGGT CTCGCTGTTC	540
60	CTTGGGAGGG TCTCCTCTGA GTGATTGACT ACCCGTCAGC GGGGGTCTTT CATTTGGGGG	600
UU		
	CTCGTCCGGG ATCGGGAGAC CCCTGCCCAG GGACCACCGA CCCACCACCG GGAGGTAAGC	660

65 CTGCGTCGGT ACTAGTTAGC TAACTAGCTC TGTATCTGGC GGACCCGTGG TGGAACTGAC 780

	GAGTTCGGAA	CACCCGGCCG	CAACCCTGGG	AGACGTCCCA	GGAGGAACAG	GGGAGGATCA	840
	GGGACGCCTG	GTGGACCCCT	TTGAAGGCCA	AGAGACCATT	TGGGGTTGCG	AGATCGTGGG	900
5	TTCGAGTCCC	ACCTCGTGCC	CAGTTGCGAG	ATCGTGGGTT	CGAGTCCCAC	CTCGTGTTTT	960
	GTTGCGAGAT	CGTGGGTTCG	AGTCCCACCT	CGCGTCTGGT	CACGGGATCG	TGGGTTCGAG	1020
10	TCCCACCTCG	TGTTTTGTTG	CGAGATCGTG	GGTTCGAGTC	CCACCTCGCG	TCTGGTCACG	1080
10	GGATCGTGGG	TTCGAGTCCC	ACCTCGTGCA	GAGGGTCTCA	ATTGGCCGGC	CTTAGAGAGG	1140
	CCATCTGATT	CTTCTGGTTT	CTCTTTTTGT	CTTAGTCTCG	TGTCCGCTCT	TGTTGTGACT	1200
15	ACTGTTTTTC	TAAAAATGGG	ACAATCTGTG	TCCACTCCCC	TTTCTCTGAC	TCTGGTTCTG	1260
	TCGCTTGGTA	ATTTTGTTTG	TTTACGTTTG	TTTTTGTGAG	TCGTCTATGT	TGTCTGTTAC	1320
20	TATCTTGTTT	TTGTTTGTGG	TTTACGGTTT	CTGTGTGTGT	CTTGTGTGTC	TCTTTGTGTT	1380
20	CAGACTTGGA	CTGATGACTG	ACGACTGTTT	TTAAGTTATG	CCTTCTAAAA	TAAGCCTAAA	1440
	AATCCTGTCA	GATCCCTATG	CTGACCACTT	CCTTTCAGAT	CAACAGCTGC	CCTTACGTAT	1500
25	CGATGGATCC	CTCGACTAAC	TAATAGCCCA	TTCTCCAAGG	TCGAGCGGGA	TCAATTCCGC	1560
	CCCCCCCTA	ACGTTACTGG	CCGAAGCCGC	TTGGAATAAG	GCCGGTGTGC	GTTTGTCTAT	1620
20	ATGTTATTTT	CCACCATATT	GCCGTCTTTT	GGCAATGTGA	GGGCCCGGAA	ACCTGGCCCT	1680
30	GTCTTCTTGA	CGAGCATTCC	TAGGGGTCTT	TCCCCTCTCG	CCAAAGGAAT	GCAAGGTCTG	1740
	TTGAATGTCG	TGAAGGAAGC	AGTTCCTCTG	GAAGCTTCTT	GAAGACAAAC	AACGTCTGTA	1800
35	GCGACCCTTT	GCAGGCAGCG	GAACCCCCCA	CCTGGCGACA	GGTGCCTCTG	CGGCCAAAAG	1860
	CCACGTGTAT	AAGATACACC	TGCAAAGGCG	GCACAACCCC	AGTGCCACGT	TGTGAGTTGG	1920
40	ATAGTTGTGG	AAAGAGTCAA	ATGGCTCTCC	TCAAGCGTAT	TCAACAAGGG	GCTGAAGGAT	1980
40	GCCCAGAAGG	TACCCCATTG	TATGGGATCT	GATCTGGGGC	CTCGGTGCAC	ATGCTTTACA	2040
	TGTGTTTAGT	CGAGGTTAAA	AAAACGTCTA	GGCCCCCGA	ACCACGGGGA	CGTGGTTTTC	2100
45	CTTTGAAAAA	CACGATAATA	ATCATGGGCG	CGGATCCCGT	CGTTTTACAA	CGTCGTGACT	2160
	GGGAAAACCC	TGGCGTTACC	CAACTTAATC	GCCTTGCAGC	ACATCCCCCT	TTCGCCAGCT	2220
50	GGCGTAATAG	CGAAGAGGCC	CGCACCGATC	GCCCTTCCCA	ACAGTTGCGC	AGCCTGAATG	2280
50	GCGAATGGCG	CTTTGCCTGG	TTTCCGGCAC	CAGAAGCGGT	GCCGGAAAGC	TGGCTGGAGT	2340
	GCGATCTTCC	TGAGGCCGAT	ACTGTCGTCG	TCCCCTCAAA	CTGGCAGATG	CACGGTTACG	2400
55	ATGCGCCCAT	CTACACCAAC	GTAACCTATC	CCATTACGGT	CAATCCGCCG	TTTGTTCCCA	2460
	CGGAGAATCC	GACGGGTTGT	TACTCGCTCA	CATTTAATGT	TGATGAAAGC	TGGCTACAGG	2520
60	AAGGCCAGAC	GCGAATTATT	TTTGATGGCG	TTAACTCGGC	GTTTCATCTG	TGGTGCAACG	2580
UU	GGCGCTGGGT	CGGTTACGGC	CAGGACAGTO	GTTTGCCGTC	TGAATTTGAC	CTGAGCGCAT	2640
	TTTTACGCGC	CGGAGAAAAC	CGCCTCGCGG	TGATGGTGCT	GCGTTGGAGT	GACGGCAGTT	2700
65	ATCTGGAAGA	TCAGGATATG	TGGCGGATGA	GCGGCATTTT	CCGTGACGTC	TCGTTGCTGC	2760

					~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	CAMCAMMECA	2820
					TCGCTTTAAT		
					GTTGCGTGAC		2880
5	TAACAGTTTC	TTTATGGCAG	GGTGAAACGC	AGGTCGCCAG	CGGCACCGCG	CCTTTCGGCG	2940
	GTGAAATTAT	CGATGAGCGT	GGTGGTTATG	CCGATCGCGT	CACACTACGT	CTGAACGTCG	3000
10	AAAACCCGAA	ACTGTGGAGC	GCCGAAATCC	CGAATCTCTA	TCGTGCGGTG	GTTGAACTGC	3060
10	ACACCGCCGA	CGGCACGCTG	ATTGAAGCAG	AAGCCTGCGA	TGTCGGTTTC	CGCGAGGTGC	3120
	GGATTGAAAA	TGGTCTGCTG	CTGCTGAACG	GCAAGCCGTT	GCTGATTCGA	GGCGTTAACC	3180
15	GTCACGAGCA	TCATCCTCTG	CATGGTCAGG	TCATGGATGA	GCAGACGATG	GTGCAGGATA	3240
	TCCTGCTGAT	GAAGCAGAAC	AACTTTAACG	CCGTGCGCTG	TTCGCATTAT	CCGAACCATC	3300
20	CGCTGTGGTA	CACGCTGTGC	GACCGCTACG	GCCTGTATGT	GGTGGATGAA	GCCAATATTG	3360
20	AAACCCACGG	CATGGTGCCA	ATGAATCGTC	TGACCGATGA	TCCGCGCTGG	CTACCGGCGA	3420
25	TGAGCGAACG GGTCGCTGGG	CGTAACGCGA GAATGAATCA	ATGGTGCAGC GGCCACGGCG	GCGATCGTAA CTAATCACGA	TCACCCGAGT CGCGCTGTAT	GTGATCATCT CGCTGGATCA	3480 3540
25	AATCTGTCGA	TCCTTCCCGC	CCGGTGCAGT	ATGAAGGCGG	CGGAGCCGAC	ACCACGGCCA	3600
	CCGATATTAT	TTGCCCGATG	TACGCGCGCG	TGGATGAAGA	CCAGCCCTTC	CCGGCTGTGC	3660
30	CGAAATGGTC	CATCAAAAAA	TGGCTTTCGC	TACCTGGAGA	GACGCGCCCG	CTGATCCTTT	3720
	GCGAATACGC	CCACGCGATG	GGTAACAGTC	TTGGCGGTTT	CGCTAAATAC	TGGCAGGCGT	3780
	TTCGTCAGTA	TCCCCGTTTA	CAGGGCGGCT	TCGTCTGGGA	CTGGGTGGAT	CAGTCGCTGA	3840
35	TTAAATATGA	TGAAAACGGC	AACCCGTGGT	CGGCTTACGG	CGGTGATTTT	GGCGATACGC	3900
	CGAACGATCG	CCAGTTCTGT	ATGAACGGTC	TGGTCTTTGC	CGACCGCACG	CCGCATCCAG	3960
40	CGCTGACGGA	AGCAAAACAC	CAGCAGCAGT	TTTTCCAGTT	CCGTTTATCC	GGGCAAACCA	4020
	TCGAAGTGAC	CAGCGAATAC	CTGTTCCGTC	ATAGCGATAA	CGAGCTCCTG	CACTGGATGG	4080
	TGGCGCTGGA	TGGTAAGCCG	CTGGCAAGCG	GTGAAGTGCC	TCTGGATGTC	GCTCCACAAG	4140
45	GTAAACAGTT	GATTGAACTG	CCTGAACTAC	CGCAGCCGGA	GAGCGCCGGG	CAACTCTGGC	4200
	TCACAGTACG	CGTAGTGCAA	CCGAACGCGA	CCGCATGGTC	AGAAGCCGGG	CACATCAGCG	4260
50	CCTGGCAGCA	GTGGCGTCTG	GCGGAAAACC	TCAGTGTGAC	GCTCCCCGCC	GCGTCCCACG	4320
	CCATCCCGCA	TCTGACCACC	AGCGAAATGG	ATTTTTGCAT	CGAGCTGGGT	AATAAGCGTT	4380
Ω.	GGCAATTTAA	CCGCCAGTCA	GGCTTTCTTT	CACAGATGTG	GATTGGCGAT	AAAAAACAAC	4440
55	TGCTGACGCC	GCTGCGCGAT	CAGTTCACCC	GTGCACCGCT	GGATAACGAC	ATTGGCGTAA	4500
	GTGAAGCGAC	CCGCATTGAC	CCTAACGCCT	GGGTCGAACG	CTGGAAGGCG	GCGGGCCATT	4560
60	ACCAGGCCGA	AGCAGCGTTG	TTGCAGTGCA	CGGCAGATAC	CACTTGCTGAT	GCGGTGCTGA	4620
	TTACGACCGC	TCACGCGTGG	CAGCATCAGG	GGAAAACCTI	ATTTATCAGC	CGGAAAACCT	4680
	ACCGGATTG	A TGGTAGTGGT	CAAATGGCGA	TTACCGTTG	A TGTTGAAGTG	GCGAGCGATA	4740
65	CACCGCATC	GGCGCGGATT	GGCCTGAACT	GCCAGCTGGG	GCAGGTAGCA	GAGCGGGTAA	4800
•							

							-
	ACTGGCTCGG	ATTAGGGCCG	CAAGAAAACT	ATCCCGACCG	CCTTACTGCC	GCCTGTTTTG	4860
	ACCGCTGGGA	TCTGCCATTG	TCAGACATGT	ATACCCCGTA	CGTCTTCCCG	AGCGAAAACG	4920
5	GTCTGCGCTG	CGGGACGCGC	GAATTGAATT	ATGGCCCACA	CCAGTGGCGC	GGCGACTTCC	4980
	AGTTCAACAT	CAGCCGCTAC	AGTCAACAGC	AACTGATGGA	AACCAGCCAT	CGCCATCTGC	5040
10	TGCACGCGGA	AGAAGGCACA	TGGCTGAATA	TCGACGGTTT	CCATATGGGG	ATTGGTGGCG	5100
	ACGACTCCTG	GAGCCCGTCA	GTATCGGCGG	AATTCCAGCT	GAGCGCCGGT	CGCTACCATT	5160
1.5	ACCAGTTGGT	CTGGTGTCAA	АААТААТААТ	AACCGGGCAG	GGGGGATCCG	AAGGCGGGGA	5220
15	CAGCAGTGCA	GTGGTGGACA	GAAAGCAAGT	GATCTAGGCC	AGCAGCCTCC	CTAAAGGGAC	5280
	TTCAGCCCAC	AAAGCCAAAC	TTGTGGCTTT	AATACAAGCT	CTGTAAATGG	TAAAAAAAA	5340
20	AAAGTCTACA	CGGACAGCAG	GTATGCTCTT	GCCACTGTAC	AGAGCAATAT	ACAGACAAAG	5400
	AGAACTGTTG CAAATCTAAC	ACATCTGCAG CGCCCAGGCA	AGAAAGACCT TCCTAAAGAG	AAGATGCTGT CAATGATCCT	GGCTAAAAGA GACAGTCTGA	AATCAGATGG AGACTATCAA	5460 5520
25	GTTATAGACA	AATTAAGACT	GGTAAAAAAA	ACCCTGTATA	AAATAGTAAA	AACTGAAAAA	5580
	AGAAAACTAG	TCCTCTCATG	AGAAGACAGA	CCTGACATCT	ACTGAAAAAT	AGACTTTACT	5640
20	GGAAAAAATA	TGTGTATGAA	TACCTTCTAG	TTTTTGTGAA	CGTTCTCAAG	ATGGATAAAA	5700
30	GCTTTTCCTT	GTAAAACGAG	ACTGATCAGA	TAGTCATCAA	GAAGATTGTT	AAAGAAAATT	5760
	TTCCAAGGTT	CGGAGTGCCA	AAAGCAATAG	TGTCAGATAA	TGGTCCTGCC	TTTGTTGCCC	5820
35	AGGTAAGTCA	GGGTGTGGCC	AAGTATTTAG	AGGTCAAATG	AAAATTCCAT	TGTGTGTACA	5880
	GACCTCAGAG	CTCAGGAAAG	ATAAAAAAGA	АТАААТАААА	CTCTAAACAG	ACCTTGACAA	5940
40	AATTAATCCT	AGAGACTGGC	ACAGACTTAC	TTGGTACTCC	TTCCCCTTGC	CCTATTTAGA	6000
40	ACTGAGAATA	CTCCCTCTTG	ATTCGGTTTT	ACTCTTTTTA	AGATCCTTTA	TGGGGCTCCT	6060
	ATGCCATCAC	TGTCTTAAAT	GATGTGTTTA	AACCTATGTT	GTTATAATAA	TGATCTATAT	6120
45	GTTAAGTTAA	AAGGCTTGCA	GGTGGTGCAG	AAAGAAGTCT	GGTCACAACT	GGCTACAGTG	6180
	AACAAGCTGG	GTACCCCAAG	GACATCTTAC	CAGTTCCAGC	CAGAGATCTG	ATCTACGATC	6240
50	CCCGGGTCGA	CCCGGGTCGA	CCCTGTGGAA	TGTGTGTCAG	TTAGGGTGTG	GAAAGTCCCC	6300
30	AGGCTCCCCA	GCAGGCAGAA	GTATGCAAAG	CATGCATCTC	AATTAGTCAG	CAACCAGGTG	6360
	TGGAAAGTCC	CCAGGCTCCC	CAGCAGGCAG	AAGTATGCAA	AGCATGCATC	TCAATTAGTC	6420
55	AGCAACCATA	GTCCCGCCCC	TAACTCCGCC	CATCCCGCCC	CTAACTCCGC	CCAGTTCCGC	6480
	CCATTCTCCG	CCCCATGGCT	GACTAATTTT	TTTTATTTAT	GCAGAGGCCG	AGGCCGCCTC	6540
60	GGCCTCTGAG	CTATTCCAGA	AGTAGTGAGG	AGGCTTTTTT	GGAGGCCTAG	GCTTTTGCAA	6600
50	AAAGCTTCAC	GCTGCCGCAA	GCACTCAGGG	CGCAAGGGCT	GCTAAAGGAA	GCGGAACACG	6660
	TAGAAAGCCA	GTCCGCAGAA	ACGGTGCTGA	CCCCGGATGA	A ATGTCAGCTA	CTGGGCTATC	6720
65	TGGACAAGGG	AAAACGCAAG	CGCAAAGAGA	AAGCAGGTAG	CTTGCAGTGG	GCTTACATGG	6780

	CGATAGCTAG	ACTGGGCGGT	TTTATGGACA	GCAAGCGAAC	CGGAATTGCC	AGCTGGGGCG	6840
	CCCTCTGGTA	AGGTTGGGAA	GCCCTGCAAA	GTAAACTGGA	TGGCTTTCTT	GCCGCCAAGG	6900
5	ATCTGATGGC	GCAGGGGATC	AAGATCTGAT	CAAGAGACAG	GATGAGGATC	GTTTCGCATG	6960
	ATTGAACAAG	ATGGATTGCA	CGCAGGTTCT	CCGGCCGCTT	GGGTGGAGAG	GCTATTCGGC	7020
10	TATGACTGGG	CACAACAGAC	AATCGGCTGC	TCTGATGCCG	CCGTGTTCCG	GCTGTCAGCG	7080
10	CAGGGGCGCC	CGGTTCTTTT	TGTCAAGACC	GACCTGTCCG	GTGCCCTGAA	TGAACTGCAG	7140
	GACGAGGCAG	CGCGGCTATC	GTGGCTGGCC	ACGACGGGCG	TTCCTTGCGC	AGCTGTGCTC	7200
15	GACGTTGTCA	CTGAAGCGGG	AAGGGACTGG	CTGCTATTGG	GCGAAGTGCC	GGGGCAGGAT	7260
	CTCCTGTCAT	CTCACCTTGC	TCCTGCCGAG	AAAGTATCCA	TCATGGCTGA	TGCAATGCGG	7320
20	CGGCTGCATA	CGCTTGATCC	GGCTACCTGC	CCATTCGACC	ACCAAGCGAA	ACATCGCATC	7380
20	GAGCGAGCAC CATCAGGGGC	GTACTCGGAT TCGCGCCAGC	GGAAGCCGGT CGAACTGTTC	CTTGTCGATC GCCAGGCTCA	AGGATGATCT AGGCGCGCAT	GGACGAAGAG GCCCGACGGC	7440 7500
25	GAGGATCTCG	TCGTGACCCA	TGGCGATGCC	TGCTTGCCGA	ATATCATGGT	GGAAAATGGC	7560
25	CGCTTTTCTG	GATTCATCGA	CTGTGGCCGG	CTGGGTGTGG	CGGACCGCTA	TCAGGACATA	7620
	GCGTTGGCTA	CCCGTGATAT	TGCTGAAGAG	CTTGGCGGCG	AATGGGCTGA	CCGCTTCCTC	7680
30	GTGCTTTACG	GTATCGCCGC	TCCCGATTCG	CAGCGCATCG	CCTTCTATCG	CCTTCTTGAC	7740
	GAGTTCTTCT	GAGCGGGACT	CTGGGGTTCG	AAATGACCGA	CCAAGCGACG	CCCAACCTGC	7800
25	CATCACGAGA	TTTCGATTCC	ACCGCCGCCT	TCTATGAAAG	GTTGGGCTTC	GGAATCGTTT	7860
35	TCCGGGACGG	AATTCGTAAT	CTGCTGCTTG	САААСААААА	AACCACCGCT	ACCAGCGGTG	7920
	GTTTGTTTGC	CGGATCAAGA	GCTACCAACT	CTTTTTCCGA	AGGTAACTGG	CTTCAGCAGA	7980
40	GCGCAGATAC	CAAATACTGT	CCTTCTAGTG	TAGCCGTAGT	TAGGCCACCA	CTTCAAGAAC	8040
	TCTGTAGCAC	CGCCTACATA	CCTCGCTCTG	CTAATCCTGT	TACCAGTGGC	TGCTGCCAGT	8100
45	GGCGATAAGT	CGTGTCTTAC	CGGGTTGGAC	TCAAGACGAT	AGTTACCGGA	TAAGGCGCAG	8160
43	CGGTCGGGCT	GAACGGGGG	TTCGTGCACA	CAGCCCAGCT	TGGAGCGAAC	GACCTACACC	8220
	GAACTGAGAT	ACCTACAGCG	TGAGCATTGA	GAAAGCGCCA	CGCTTCCCGA	AGGGAGAAAG	8280
50	GCGGACAGGT	ATCCGGTAAG	CGGCAGGGTC	GGAACAGGAG	AGCGCACGAG	GGAGCTTCCA	8340
	GGGGGAAACG	CCTGGTATCT	TTATAGTCCT	GTCGGGTTTC	GCCACCTCTG	ACTTGAGCGT	8400
55	CGATTTTTGT	GATGCTCGTC	AGGGGGGGGG	AGCCTATGGA	AAAACGCCAG	CAACGCCGAG	8460
22	ATGCGCCGCC	TCGAGAACCC	TGGCCCTATT	ATTGGGTGGA	CTAACCATGG	GGGGAATTGC	8520
	CGCTGGAATA	GGAACAGGGA	CTACTGCTCT	AATGGCCACT	CAGCAATTCC	AGCAGCTCCA	8580
60	AGCCGCAGTA	CAGGATGATC	TCAGGGAGGT	TGAAAAATCA	ATCTCTAACC	TAGAAAAGTC	8640
	TCTCACTTCC	CTGTCTGAAG	TTGTCCTACA	GAATCGAAGG	GGCCTAGACT	TGTTATTTCT	8700
65	AAAAGAAGGA	GGGCTGTGTG	CTGCTCTAAA	AGAAGAATGT	TGCTTCTATG	CGGACCACAC	8760
05	AGGACTAGTG	AGAGACAGCA	TGGCCAAATT	GAGAGAGAGG	CTTAATCAGA	GACAGAAACT	8820

	GTTTGAGTCA ACTCAAGGAT GGTTTGAGGG ACTGTTTAAC AGATCCCCTT GGTTTACCAC	8880
	CTTGATATCT ACCATTATGG GACCCCTCAT TGTACTCCTA ATGATTTTGC TCTTCGGACC	8940
5	CTGCATTCTT AATCGATTAG TCCAATTTGT TAAAGACAGG ATATCAGTGG TCCAGGCTCT	9000
	AGTTTTGACT CAACAATATC ACCAGCTGAA GCCTATAGAG TACGAGCCAT AGATAAAATA	9060
10	AAAGATTTTA TTTAGTCTCC AGAAAAAGGG GGG	9093
	(2) INFORMATION FOR SEQ ID NO:24:	
15	(i) SEQUENCE CHARACTERISTICS:  (A) LENGTH: 46 base pairs  (B) TYPE: nucleic acid  (C) STRANDEDNESS: single  (D) TOPOLOGY: linear	
20	(ii) MOLECULE TYPE: DNA (genomic)	
25	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:24:	
23	GACTAACCTT GATTCCCTGG AGGCGGGGGT CTTTCATTTG GGGGCT	46
	(2) INFORMATION FOR SEQ ID NO:25:	
30	<ul><li>(i) SEQUENCE CHARACTERISTICS:</li><li>(A) LENGTH: 4834 base pairs</li><li>(B) TYPE: nucleic acid</li></ul>	
2.5	(C) STRANDEDNESS: single (D) TOPOLOGY: linear	
35	(ii) MOLECULE TYPE: DNA (genomic)	
40	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:25:	
	TGAAGAATAA AAAATTACTG GCCTCTTGTG AGAACATGAA CTTTCACCTC GGAGCCCACC	60
45	CCCTCCCATC TGGAAAACAT ACTTGAGAAA AACATTTTCT GGAACAACCA CAGAATGTTT	120
1,5	CAACAGGCCA GATGTATTGC CAAACACAGG ATATGACTCT TTGGTTGAGT AAATTTGTGG	180
	TTGTTAAACT TCCCCTATTC CCTCCCCATT CCCCCTCCCA GTTTGTGGTT TTTTCCTTTA	240
50	AAAGCTTGTG AAAAATTTGA GTCGTCGTCG AGACTCCTCT ACCCTGTGCA AAGGTGTATG	300
	AGTTTCGACC CCAGAGCTCT GTGTGCTTTC TGTTGCTGCT TTATTTCGAC CCCAGAGCTC	360
55	TGGTCTGTGT GCTTTCATGT CGCTGCTTTA TTAAATCTTA CCTTCTACAT TTTATGTATG	420
	GTCTCAGTGT CTTCTTGGGT ACGCGGCTGT CCCGGGACTT GAGTGTCTGA GTGAGGGTCT	480
	TCCCTCGAGG GTCTTTCATT TGGTACATGG GCCGGGAATT CGAGAATCTT TCATTTGGTG	540
60	CATTGGCCGG GAATTCGAAA ATCTTTCATT TGGTGCATTG GCCGGGAAAC AGCGCGACCA	600
	CCCAGAGGTC CTAGACCCAC TTAGAGGTAA GATTCTTTGT TCTGTTTTGG TCTGATGTCT	660
65	GTGTTCTGAT GTCTGTTTC TGTTTCTAAG TCTGGTGCGA TCGCAGTTTC AGTTTTGCGG	720

	ACGCTCAGTG	AGACCGCGCT	CCGAGAGGGA	GTGCGGGGTG	GATAAGGATA	GACGTGTCCA	780
•	GGTGTCCACC	GTCCGTTCGC	CCTGGGAGAC	GTCCCAGGAG	GAACAGGGGA	GGATCAGGGA	840
5	CGCCTGGTGG	ACCCCTTTGA	AGGCCAAGAG	ACCATTTGGG	GTTGCGAGAT	CGTGGGTTCG	900
	AGTCCCACCT	CGTGCCCAGT	TGCGAGATCG	TGGGTTCGAG	TCCCACCTCG	TGTTTTGTTG	960
10	CGAGATCGTG	GGTTCGAGTC	CCACCTCGCG	TCTGGTCACG	GGATCGTGGG	TTCGAGTCCC	1020
10	ACCTCGTGTT	TTGTTGCGAG	ATCGTGGGTT	CGAGTCCCAC	CTCGCGTCTG	GTCACGGGAT	1080
	CGTGGGTTCG	AGTCCCACCT	CGTGCAGAGG	GTCTCAATTG	GCCGGCCTTA	GAGAGGCCAT	1140
15	CTGATTCTTC	TGGTTTCTCT	TTTTGTCTTA	GTCTCGTGTC	CGCTCTTGTT	GTGACTACTG	1200
	TTTTTCTAAA	AATGGGACAA	TCTGTGTCCA	CTCCCCTTTC	TCTGACTCTG	GTTCTGTCGC	1260
20	TTGGTAATTT TTGTTTTTGT		CGTTTGTTTT CGGTTTCTGT	TGTGAGTCGT GTGTGTCTTG	CTATGTTGTC TGTGTCTCTT		1320 1380
	CTTGGACTGA	TGACTGACGA	CTGTTTTTAA	GTTATGCCTT	CTAAAATAAG	CCTAAAAATC	1440
25	CTGTCAGATC	CCTATGCTGA	CCACTTCCTT	TCAGATCAAC	AGCTGCCCTG	CCTCCCACTC	1500
25	CAACTCCAGA	GAGCAGCCAG	CGGGTCACAG	TGGTCCCGCC	CATGAACCTG	GAGCCTAGGG	1560
	AAAAATGAGC	TCGGAAATCC	GGAGCAAATG	AGGAGTGGTC	CCTGAGAAGT	CAGTGGCCTA	1620
30	AATGTTGTGG	CTGCTGAAGC	AAAAGAAGAG	GAGGCTGTTC	GAGTAGCCGG	CCAAGAGCGC	1680
	CGCGGGTTCC	CAGGCAGCTT	CTCATTCCCC	TGTCCCTCCC	ATCCCGTCTC	TTGTTAACAG	1740
25	AAAAACTGCT	TTCACTTTGA	GATATGAGTG	GCCCGATACA	GCCAGCTGTG	AGAGCTGTAC	1800
35	TCCCTTCCCT	GCCCCACGTG	TTTTCTCTTC	TCAGGCGACC	CCTCCCTGAG	CTGCTGGCAG	1860
	TGAGTCTGTT	CTAAGCTCCA	GTGAGGGAGG	CATCCGCCCA	CTTGGGGCTT	CTGTCCAAGG	1920
40	TAAGGAGCAC	CTGTGAGTCT	AACTGCCAGG	CTCTGATGGG	GGTCTCGTCT	CTGTGGGACT	1980
	AGAAAGTGTC	CCAACAATCT	GACCAAGGTA	ACAGGAAGTT	AAGACAAAGA	CAGAGACCAA	2040
45	AGTCAGAATC	AGAGCTGTGC	TGTGAGACAA	AAAGATAAAA	. <b>ААААТААА</b> А	GCTGGCCACA	2100
45	AAAGTCAGGA	AAACTAGAAA	ACTTAGATAG	TACCTGGCAA	CAAAAGAAAG	CTTTTGGCTA	2160
	AAGATCAACG	TGTATACTGT	AAAGAAAATG	AGCACTGGGT	GAGAGACTGC	CCCAACAAAA	2220
50	AGAAGAGGAG	CCCCCTCAT	GACCAAACCC	TTCACCTGTT	CGTGGCTAAA	AGTAAAGAGA	2280
	TAACAAAAGG	GGTGCTAACA	CAGAAGCTGA	GTCCTTAAAA	GAGTCCGGTG	GCCTACCTGT	2340
~ ~	TGAAGCAGCT	AAAAAAGAGA	CTGTGTTTCA	TACTCCTCCA	CTGACCAGTG	CAAAACAAGC	2400
55	TAAAAAGTTC	CTGGGCACTG	CGGGCTTTTG	CAGATTGTGG	ATTCCAGGTT	TTGCTGAGTT	2460
	AAAGAGATAA	ACAGCCCTTC	GTATAGAAAA	ATAAAAAAC <i>I</i>	A ACCTTGGATG	TCCTTGGATG	2520
60	CTATTGAGAC	TGCCCTAATG	TTGTCCCCAG	CTATGGGACT	CCTAGATGTG	ACTGAGAACA	2580
	AAGGTATTGC	CAAAGAAGTT	CTTACTCAGA	GATTGGGAC	CTGAAAAAGA	CCTGTGGCAT	2640
(=	ACTTGTAAGA	AATTAGACCT	GGTGGCTGTA	AGATGGCCT	G CTTGTCTGCA	CATAGTGGCT	2700
65	TCTGGTCAAG	GACGCAGAT <i>I</i>	A AATTGACTCT	GAGACAAAA	TTGGCACATG	TCCTAGAAAG	2760

	TGTGGTTCAG	CCCCCATGAC	CGATGGCTGA	CTAACGCTCT	TGAAAACATT	ATCCAACTGT	2820
	TCCCCTGACC	GATGGACACA	TTGTCAGAGC	TTTTTTTGAC	TGAACGAGTG	ACCTTCGCTC	2880
5	CCCCTGCTAT	CCTCGATCTC	ACTACTGCCT	GAGACTTCAC	CTACTCATCA	TTGTGCTGAC	2940
	ATTCTGGCAG	AAGAAACTCA	TACTCGAAAT	GATCTGAAGG	ATCAGATCAG	CCTTGGCCTG	3000
10	AGAGTTTGAG	CTGGTACACG	GATGGCAGTA	GCCTGGAGGT	TAAGGGTAAG	CGGAAGGCGG	3060
	GGACAGCAGT	GCAGTGGTGG	ACAGAAAGCA	AGTGATCTAG	GCCAGCAGCC	TCCCTAAAGG	3120
	GACTTCAGCC	CACAAAGCCA	AACTTGTGGC	TTTAATACAA	GCTCTGTAAA	TGGTAAAAAA	3180
15	AAAAAAGTCT	ACACGGACAG	CAGGTATGCT	CTTGCCACTG	TACAGAGCAA	TATACAGACA	3240
20	AAGAGAACTG TGGCAAATCT	TTGACATCTG AACCGCCCAG	CAGAGAAAGA GCATCCTAAA	CCTAAGATGC GAGCAATGAT	TGTGGCTAAA CCTGACAGTC	AGAAATCAGA TGAAGACTAT	3300 3360
20	CAAGTTATAG	ACAAATTAAG	ACTGGTAAAA	AAAACCCTGT	ATAAAATAGT	AAAAACTGAA	3420
	AAAAGAAAAC	TAGTCCTCTC	ATGAGAAGAC	AGACCTGACA	TCTACTGAAA	AATAGACTTT	3480
25	ACTGGAAAAA	ATATGTGTAT	GAATACCTTC	TAGTTTTTGT	GAACGTTCTC	AAGATGGATA	3540
	AAAGCTTTTC	CTTGTAAAAC	GAGACTGATC	AGATAGTCAT	CAAGAAGATT	GTTAAAGAAA	3600
20	ATTTTCCAAG	GTTCGGAGTG	CCAAAAGCAA	TAGTGTCAGA	TAATGGTCCT	GCCTTTGTTG	3660
. 30	CCCAGGTAAG	TCAGGGTGTG	GCCAAGTATT	TAGAGGTCAA	ATGAAAATTC	CATTGTGTGT	3720
	ACAGACCTCA	GAGCTCAGGA	AAGATAAAAA	AGAATAAATA	AAACTCTAAA	CAGACCTTGA	3780
35	CAAAATTAAT	CCTAGAGACT	GGCACAGACT	TACTTGGTAC	TCCTTCCCCT	TGCCCTATTT	3840
	AGAACTGAGA	ATACTCCCTC	TTGATTCGGT	TTTACTCTTT	TTAAGATCCT	TTATGGGGCT	3900
40	CCTATGCCAT	CACTGTCTTA	AATGATGTGT	TTAAACCTAT	GTTGTTATAA	TAATGATCTA	3960
40	TATGTTAAGT	TAAAAGGCTT	GCAGGTGGTG	CAGAAAGAAG	TCTGGTCACA	ACTGGCTACA	4020
	GTGAACAAGC	TGGGTACCCC	AAGGACATCT	TACCAGTTCC	AGCCAGAGAT	CTGATCTACG	4080
45	TACACCTGCG	TCATGCTGAG	ACCCTCAAGC	CTCACTAAAA	GGGTCCCTGC	CTAGTTCTGT	4140
	TTACTAATCT	GCCTTATTCT	GTTTTTGTTC	CCATGTTAAA	GATAGAGTAA	ATGCAGTATT	4200
50	CTCCACATAG	AGATATAGAC	TTCTGAAATT	CTAAGATTAG	AATTATTTAC	AAGAAGAAGT	4260
50	GGGGAATGAA	GAATAAAAAA	TTACTGGCCT	CTTGTGAGAA	CATGAACTTT	CACCTCGGAG	4,320
	CCCACCCCT	CCCATCTGGA	AAACATACTT	GAGAAAAACA	TTTTCTGGAA	CAACCACAGA	4380
55	ATGTTTCAAC	AGGCCAGATG	TATTGCCAAA	. CACAGGATAI	GACTCTTTGG	TTGAGTAAAT	4440
	TTGTGGTTGT	TAAACTTCCC	: CTATTCCCTC	CCCATTCCC	CTCCCAGTTT	GTGGTTTTTT	4500
60	CCTTTAAAAG	CTTGTGAAAA	ATTTGAGTCG	TCGTCGAGAC	TCCTCTACCC	TGTGCAAAGG	4560
υU	TGTATGAGTT	TCGACCCCAG	AGCTCTGTGT	GCTTTCTGTT	GCTGCTTTAT	TTCGACCCCA	4620
	GAGCTCTGGT	CTGTGTGCT1	TCATGTCGCT	GCTTTATTA	ATCTTACCTT	CTACATTTTA	4680
65	TGTATGGTCT	CAGTGTCTTC	TTGGGTACGC	GGCTGTCCC	GGACTTGAGT	GTCTGAGTGA	4740

PCT/US98/03918 WO 98/38326 104 GGGTCTTCCC TCGAGGGTCT TTCATTTGGT ACATGGGCCG GGAATTCGAG AATCTTTCAT 4800 4834 TTGGTGCATT GGCCGGGAAT TCGAAAATCT TTCA (2) INFORMATION FOR SEQ ID NO:26: 5 (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 4518 base pairs (B) TYPE: nucleic acid 10 (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: DNA (genomic) 15 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:26: CACCTGACGC GCCCTGTAGC GGCGCATTAA GCGCGGCGGG TGTGGTGGTT ACGCGCAGCG 60 20 TGACCGCTAC ACTTGCCAGC GCCCTAGCGC CCGCTCCTTT CGCTTTCTTC CCTTCCTTTC 120 TCGCCACGTT CGCCGGCTTT CCCCGTCAAG CTCTAAATCG GGGGCTCCCT TTAGGGTTCC 180 25 GATTTAGTGC TTTACGGCAC CTCGACCCCA AAAAACTTGA TTAGGGTGAT GGTTCACGTA 240 GTGGGCCATC GCCCTGATAG ACGGTTTTTC GCCCTTTGAC GTTGGAGTCC ACGTTCTTTA 300 ATAGTGGACT CTTGTTCCAA ACTGGAACAA CACTCAACCC TATCTCGGTC TATTCTTTTG 360 30 420 ATTTATAAGG GATTTTGCCG ATTTCGGCCT ATTGGTTAAA AAATGAGCTG ATTTAACAAA AATTTAACGC GAATTTTAAC AAAATATTAA CGCTTACAAT TTACGCGTTA AGATACATTG 480 35 ATGAGTTTGG ACAAACCACA ACTAGAATGC AGTGAAAAAA ATGCTTTATT TGTGAAATTT 540 GTGATGCTAT TGCTTTATTT GTAACCATTA TAAGCTGCAA TAAACAAGTT AACAACAACA 600 ATTGCATTCA TTTTATGTTT CAGGTTCAGG GGGAGGTGTG GGAGGTTTTT TAAAGCAAGT 660 40 AAAACCTCTA CAAATGTGGT ATGGCTGATT ATGATCATGA ACAGACTGTG AGGACTGAGG 720 GGCCTGAAAT GAGCCTTGGG ACTGTGAATC TAAAATACAC AAACAATTAG AATCAGTAGT 780 45 TTAACACATT ATACACTTAA AAATTGGATC TCCATTCGCC ATTCAGGCTG CGCAACTGTT 840 GGGAAGGGCG ATCGGTGCGG GCCTCTTCGC TATTACGCCA GCTGGCGAAA GGGGGATGTG 900 CTGCAAGGCG ATTAAGTTGG GTAACGCCAG GGTTTTCCCA GTCACGACGT TGTAAAACGA 960 50

CGGCCAGTGA ATTGTAATAC GACTCACTAT AGGGCGAATT GGGTACACTT ACCTGGTACC

CCACCCGGGT GGAAAATCGA TGGGCCCGCG GCCGCTCTAG AAGTACTCTC GAGAAGCTTT

TTGAATTCTT TGGATCCACT AGTGTCGACC TGCAGGCGCG CGAGCTCCAG CTTTTGTTCC

CTTTAGTGAG GGTTAATTTC GAGCTTGGCG TAATCAAGGT CATAGCTGTT TCCTGTGTGA

AATTGTTATC CGCTCACAAT TCCACACAAT ATACGAGCCG GAAGTATAAA GTGTAAAGCC

TGGGGTGCCT AATGAGTGAG CTAACTCACA GTAATTGCGG CTAGCGGATC TGACGGTTCA

CTARACCAGC TCTGCTTATA TAGACCTCCC ACCGTACACG CCTACCGCCC ATTTGCGTCA

ATGGGGCGGA GTTGTTACGA CATTTTGGAA AGTCCCGTTG ATTTTGGTGC CAAAACAAAC

55

60

65

1020

1080

1140

1200

1260

1320

1380

1440

							-
	TCCCATTGAC	GTCAATGGGG	TGGAGACTTG	GAAATCCCCG	TGAGTCAAAC	CGCTATCCAC	1500
5	GCCCATTGAT	GTACTGCCAA	AACCGCATCA	CCATGGTAAT	AGCGATGACT	AATACGTAGA	1560
3	TGTACTGCCA	AGTAGGAAAG	TCCCATAAGG	TCATGTACTG	GGCATAATGC	CAGGCGGGCC	1620
	ATTTACCGTC	ATTGACGTCA	ATAGGGGGCG	TACTTGGCAT	ATGATACACT	TGATGTACTG	1680
10	CCAAGTGGGC	AGTTTACCGT	AAATACTCCA	CCCATTGACG	TCAATGGAAA	GTCCCTATTG	1740
	GCGTTACTAT	GGGAACATAC	GTCATTATTG	ACGTCAATGG	GCGGGGGTCG	TTGGGCGGTC	1800
15	AGCCAGGCGG	GCCATTTACC	GTAAGTTATG	TAACGCGGAA	CTCCATATAT	GGGCTATGAA	1860
13					TGGCGGTAAT AAGGCCAGCA		1920 1980
20	AACCGTAAAA	AGGCCGCGTT	GCTGGCGTTT	TTCCATAGGC	TCCGCCCCCC	TGACGAGCAT	2040
20	CACAAAAATC	GACGCTCAAG	TCAGAGGTGG	CGAAACCCGA	CAGGACTATA	AAGATACCAG	2100
	GCGTTTCCCC	CTGGAAGCTC	CCTCGTGCGC	TCTCCTGTTC	CGACCCTGCC	GCTTACCGGA	2160
25	TACCTGTCCG	CCTTTCTCCC	TTCGGGAAGC	GTGGCGCTTT	CTCATAGCTC	ACGCTGTAGG	2220
	TATCTCAGTT	CGGTGTAGGT	CGTTCGCTCC	AAGCTGGGCT	GTGTGCACGA	ACCCCCGTT	2280
30	CAGCCCGACC	GCTGCGCCTT	ATCCGGTAAC	TATCGTCTTG	AGTCCAACCC	GGTAAGACAC	2340
50	GACTTATCGC	CACTGGCAGC	AGCCACTGGT	AACAGGATTA	GCAGAGCGAG	GTATGTAGGC	2400
	GGTGCTACAG	AGTTCTTGAA	GTGGTGGCCT	AACTACGGCT	ACACTAGAAG	GACAGTATTT	2460
35	GGTATCTGCG	CTCTGCTGAA	GCCAGTTACC	TTCGGAAAAA	GAGTTGGTAG	CTCTTGATCC	2520
	GGCAAACAAA	CCACCGCTGG	TAGCGGTGGT	TTTTTTGTTT	GCAAGCAGCA	GATTACGCGC	2580
40	AGAAAAAAAG	GATCTCAAGA	AGATCCTTTG	ATCTTTTCTA	CGGGGTCTGA	CGCTCAGTGG	2640
40	AACGAAAACT	CACGTTAAGG	GATTTTGGTC	ATGAGATTAT	CAAAAAGGAT	CTTCACCTAG	2700
	ATCCTTTTAA	ATTAAAAATG	AAGTTTTAAA	TCAATCTAAA	GTATATATGA	GTAACCTGAG	2760
45	GCTATGGCAG	GGCCTGCCGC	CCCGACGTTG	GCTGCGAGCC	CTGGGCCTTC	ACCCGAACTT	2820
	GGGGGGTGGG	GTGGGGAAAA	GGAAGAAACG	CGGGCGTATT	GGCCCCAATG	GGGTCTCGGT	2880
50	GGGGTATCGA	CAGAGTGCCA	GCCCTGGGAC	CGAACCCCGC	GTTTATGAAC	AAACGACCCA	2940
50	ACACCGTGCG	TTTTATTCTG	TCTTTTTATT	GCCGTCATAG	CGCGGGTTCC	TTCCGGTATT	3000
	GTCTCCTTCC	GTGTTTCAGT	TAGCCTCCCC	CTAGGGTGGG	CGAAGAACTC	CAGCATGAGA	3060
55	TCCCCGCGCT	GGAGGATCAT	CCAGCCGGCG	TCCCGGAAAA	CGATTCCGAA	GCCCAACCTT	3120
	TCATAGAAGG	CGGCGGTGGA	ATCGAAATCT	CGTGATGGCA	GGTTGGGCGT	CGCTTGGTCG	3180
60	GTCATTTCGA	ACCCCAGAGT	CCCGCTCAGA	AGAACTCGTC	AAGAAGGCGA	TAGAAGGCGA	3240
00	TGCGCTGCGA	ATCGGGAGCG	GCGATACCGT	AAAGCACGAG	GAAGCGGTCA	GCCCATTCGC	3300
	CGCCAAGCTC	TTCAGCAATA	TCACGGGTAG	CCAACGCTAT	GTCCTGATAG	CGGTCCGCCA	3360
65	CACCCAGCCG	GCCACAGTCG	ATGAATCCAG	AAAAGCGGCC	ATTTTCCACC	ATGATATTCG	3420

	GCAAGCAGGC ATCGCCATGG GTCACGACGA GATCCTCGCC GTCGGGCATG CTCGCCTTGA	3480
	GCCTGGCGAA CAGTTCGGCT GGCGCGAGCC CCTGATGCTC TTCGTCCAGA TCATCCTGAT	3540
5	CGACAAGACC GGCTTCCATC CGAGTACGTG CTCGCTCGAT GCGATGTTTC GCTTGGTGGT	3600
	CGAATGGGCA GGTAGCCGGA TCAAGCGTAT GCAGCCGCCG CATTGCATCA GCCATGATGG	3660
10	ATACTTTCTC GGCAGGAGCA AGGTGAGATG ACAGGAGATC CTGCCCCGGC ACTTCGCCCA	3720
10	ATAGCAGCCA GTCCCTTCCC GCTTCAGTGA CAACGTCGAG CACAGCTGCG CAAGGAACGC	3780
	CCGTCGTGGC CAGCCACGAT AGCCGCGCTG CCTCGTCTTG CAGTTCATTC AGGGCACCGG	3840
15	ACAGGTCGGT CTTGACAAAA AGAACCGGGC GCCCCTGCGC TGACAGCCGG AACACGGCGG CATCAGAGCA GCCGATTGTC TGTTGTGCCC AGTCATAGCC GAATAGCCTC TCCACCCAAG	3900 3960
	CGGCCGGAGA ACCTGCGTGC AATCCATCTT GTTCAATCAT GCGAAACGAT CCTCATCCTG	4020
20	TCTCTTGATC GATCTTTGCA AAAGCCTAGG CCTCCAAAAA AGCCTCCTCA CTACTTCTGG	4080
	AATAGCTCAG AGGCCGAGGC GGCCTCGGCC TCTGCATAAA TAAAAAAAAT TAGTCAGCCA	4140
25	TGGGGCGGAG AATGGGCGGA ACTGGGCGGA GTTAGGGGCG GGATGGGCGG AGTTAGGGGC	4200
25	GGGACTATGG TTGCTGACTA ATTGAGATGC ATGCTTTGCA TACTTCTGCC TGCTGGGGAG	4260
	CCTGGGGACT TTCCACACCT GGTTGCTGAC TAATTGAGAT GCATGCTTTG CATACTTCTG	4320
30	CCTGCTGGGG AGCCTGGGGA CTTTCCACAC CCTAACTGAC ACACATTCCA CAGCTGGTTC	4380
	TTTCCGCCTC AGGACTCTTC CTTTTTCAAT ATTATTGAAG CATTTATCAG GGTTATTGTC	4440
35	TCATGAGCGG ATACATATTT GAATGTATTT AGAAAAATAA ACAAATAGGG GTTCCGCGCA	4500
33	CATTTCCCCG AAAAGTGC	4518
	(2) INFORMATION FOR SEQ ID NO:27:	
40	(i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 25 base pairs	
	(B) TYPE: nucleic acid (C) STRANDEDNESS: single	
45	(D) TOPOLOGY: linear	
13	(ii) MOLECULE TYPE: DNA (genomic)	
50		
30	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:27:	
	CTCCACATAG AGATATAGAC TTCTG	25
55	(2) INFORMATION FOR SEQ ID NO:28:	
60	<ul> <li>(i) SEQUENCE CHARACTERISTICS:</li> <li>(A) LENGTH: 45 base pairs</li> <li>(B) TYPE: nucleic acid</li> <li>(C) STRANDEDNESS: single</li> <li>(D) TOPOLOGY: linear</li> </ul>	

(ii) MOLECULE TYPE: DNA (genomic)

65

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:28: 45 CGATCTTATT AATTAACTGG AGTTTTGAGC CCRMCCCCTC CCATC 5 (2) INFORMATION FOR SEQ ID NO:29: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 5594 base pairs (B) TYPE: nucleic acid 10 (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: DNA (genomic) 15 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:29: TGCATTAGTT ATTAATAGTA ATCAATTACG GGGTCATTAG TTCATAGCCC ATATATGGAG 60 20 TTCCGCGTTA CATAACTTAC GGTAAATGGC CCGCCTGGCT GACCGCCCAA CGACCCCCGC 120 180 CCATTGACGT CAATAATGAC GTATGTTCCC ATAGTAACGC CAATAGGGAC TTTCCATTGA 25 CGTCAATGGG TGGAGTATTT ACGGTAAACT GCCCACTTGG CAGTACATCA AGTGTATCAT 240 ATGCCAAGTA CGCCCCCTAT TGACGTCAAT GACGGTAAAT GGCCCGCCTG GCATTATGCC 300 CAGTACATGA CCTTATGGGA CTTTCCTACT TGGCAGTACA TCTACGTATT AGTCATCGCT 360 30 ATTACCATGG TGATGCGGTT TTGGCAGTAC ATCAATGGGC GTGGATAGCG GTTTGACTCA 420 CGGGGATTTC CAAGTCTCCA CCCCATTGAC GTCAATGGGA GTTTGTTTTG GCACCAAAAT 480 35 CAACGGGACT TTCCAAAATG TCGTAACAAC TCCGCCCCAT TGACGCAAAT GGGCGGTAGG 540 CGTGTACGGT GGGAGGTCTA TATAAGCAGA GCTGGTTTAG TGAACCGTCA GATCCGCGCC 600 AGTCCTCCGA TTGACTGAGT CGCCCGGGTA CCCGTGTATC CAATAAACCC TCTTGCAGTT 660 40 GCATCCGACT TGTGGTCTCG CTGTTCCTTG GGAGGGTCTC CTCTGAGTGA TTGACTACCC 720 780 GTCAGCGGGG GTCTTTCATT TGGGGGCTCG TCCGGGATCG GGAGACCCCT GCCCAGGGAC 45 840 CACCGACCCA CCACCGGGAG GTAAGCTGGC CAGCAACTTA TCTGTGTCTG TCCGATTGTC TAGTGTCTAT GACTGATTTT ATGCGCCTGC GTCGGTACTA GTTAGCTAAC TAGCTCTGTA 900 TCTGGCGGAC CCGTGGTGGA ACTGACGAGT TCGGAACACC CGGCCGCAAC CCTGGGAGAC 960 50 GTCCCAGGAG GAACAGGGGA GGATCAGGGA CGCCTGGTGG ACCCCTTTGA AGGCCAAGAG 1020 ACCATTTGGG GTTGCGAGAT CGTGGGTTCG AGTCCCACCT CGTGCCCAGT TGCGAGATCG 1080 55 TGGGTTCGAG TCCCACCTCG TGTTTTGTTG CGAGATCGTG GGTTCGAGTC CCACCTCGCG 1140 TCTGGTCACG GGATCGTGGG TTCGAGTCCC ACCTCGTGTT TTGTTGCGAG ATCGTGGGTT 1200 CGAGTCCCAC CTCGCGTCTG GTCACGGGAT CGTGGGTTCG AGTCCCACCT CGTGCAGAGG 1260 60 GTCTCAATTG GCCGGCCTTA GAGAGGCCAT CTGATTCTTC TGGTTTCTCT TTTTGTCTTA 1320 GTCTCGTGTC CGCTCTTGTT GTGACTACTG TTTTTCTAAA AATGGGACAA TCTGTGTCCA 1380 65 CTCCCCTTTC TCTGACTCTG GTTCTGTCGC TTGGTAATTT TGTTTGTTTA CGTTTGTTTT 1440

							_
	TGTGAGTCGT	CTATGTTGTC	TGTTACTATC	TTGTTTTTGT	TTGTGGTTTA	CGGTTTCTGT	1500
_	GTGTGTCTTG	TGTGTCTCTT	TGTGTTCAGA	CTTGGACTGA	TGACTGACGA	CTGTTTTTAA	1560
5	GTTATGCCTT	CTAAAATAAG	CCTAAAAATC	CTGTCAGATC	CCTATGCTGA	CCACTTCCTT	1620
	TCAGATCAAC	AGCTGCCCTT	ACGTATCGAT	GGATCCCTCG	ACTAACTAAT	AGCCCATTCT	1680
10	CCAAGGTCGA	GCGGGATCAA	TTCCGCCCCC	CCCCTAACGT	TACTGGCCGA	AGCCGCTTGG	1740
	AATAAGGCCG	GTGTGCGTTT	GTCTATATGT	TATTTTCCAC	CATATTGCCG	TCTTTTGGCA	1800
1.5	ATGTGAGGGC	CCGGAAACCT	GGCCCTGTCT	TCTTGACGAG	CATTCCTAGG	GGTCTTTCCC	1860
15	CTCTCGCCAA	AGGAATGCAA	GGTCTGTTGA	ATGTCGTGAA	GGAAGCAGTT	CCTCTGGAAG	1920
	CTTCTTGAAG	ACAAACAACG	TCTGTAGCGA	CCCTTTGCAG	GCAGCGGAAC	CCCCCACCTG	1980
20	GCGACAGGTG	CCTCTGCGGC	CAAAAGCCAC	GTGTATAAGA	TACACCTGCA	AAGGCGGCAC	2040
	AACCCCAGTG	CCACGTTGTG	AGTTGGATAG	TTGTGGAAAG	AGTCAAATGG	CTCTCCTCAA	2100
25	GCGTATTCAA	CAAGGGGCTG	AAGGATGCCC	AGAAGGTACC	CCATTGTATG	GGATCTGATC	2160
25	TGGGGCCTCG	GTGCACATGC	TTTACATGTG	TTTAGTCGAG	GTTAAAAAAA	CGTCTAGGCC	2220
	CCCCGAACCA	CGGGGACGTG	GTTTTCCTTT	GAAAAACACG	ATAATAATCA	TGGCTACAGG	2280
30	CTCCCGGACG	TCCCTGCTCC	TGGCTTTTGG	CCTGCTCTGC	CTGCCCTGGC	TTCAAGAGGG	2340
	CAGTGCCTTC	CCAACCATTC	CCTTATCCAG	GCTTTTTGAC	AACGCTATGC	TCCGCGCCCA	2400
25	TCGTCTGCAC	CAGCTGGCCT	TTGACACCTA	CCAGGAGTTT	GAAGAAGCCT	ATATCCCAAA	2460
35	GGAACAGAAG	TATTCATTCC	TGCAGAACCC	CCAGACCTCC	CTCTGTTTCT	CAGAGTCTAT	2520
	TCCGACACCC	TCCAACAGGG	AGGAAACACA	ACAGAAATCO	AACCTAGAGC	TGCTCCGCAT	2580
40	CTCCCTGCTG	CTCATCCAGT	CGTGGCTGGA	GCCCGTGCAG	TTCCTCAGGA	GTGTCTTCGC	2640
	CAACAGCCTG	GTGTACGGCG	CCTCTGACAG	CAACGTCTAT	GACCTCCTAA	AGGACCTAGA	2700
45	GGAAGGCATC	CAAACGCTGA	TGGGGAGGCT	GGAAGATGG	AGCCCCCGGA	CTGGGCAGAT	2760
43	CTTCAAGCAG	ACCTACAGCA	AGTTCGACAC	AAACTCACAC	AACGATGACG	CACTACTCAA	2820
	GAACTACGGG	CTGCTCTACT	GCTTCAGGAA	GGACATGGA	C AAGGTCGAGA	CATTCCTGCG	2880
50	CATCGTGCAG	TGCCGCTCTG	TGGAGGGCAG	CTGTGGCTT	TAGCTGCCCG	GGTGGCATCC	2940
	TGTGACCCCT	CCCCAGTGCC	TCTCCTGGCC	CTGGAAGTT	G CCACTCCAGI	GCCCACCAGC	3000
55	CTTGTCCTAA	TGTGTGTCAG	TTAGGGTGTG	GAAAGTCCC	C AGGCTCCCCA	GCAGGCAGAA	3060
))	GTATGCAAAG	CATGCATCT	AATTAGTCAG	CAACCAGGT	G TGGAAAGTCC	CCAGGCTCCC	3120
	CAGCAGGCAG	AAGTATGCA	A AGCATGCATO	TCAATTAGT	C AGCAACCATA	GTCCCGCCCC	3180
60	TAACTCCGCC	CATCCCGCC	CTAACTCCG	CCAGTTCCG	C CCATTCTCC	CCCCATGGCT	3240
						CTATTCCAGA	3300
65	••					GCTGCCGCAA	3360
05	GCACTCAGG	G CGCAAGGGC	r gctaaagga	A GCGGAACAC	G TAGAAAGCC	A GTCCGCAGAA	3420

	ACGGTGCTGA	CCCCGGATGA	ATGTCAGCTA	CTGGGCTATC	TGGACAAGGG	AAAACGCAAG	3480
	CGCAAAGAGA	AAGCAGGTAG	CTTGCAGTGG	GCTTACATGG	CGATAGCTAG	ACTGGGCGGT	3540
5	TTTATGGACA	GCAAGCGAAC	CGGAATTGCC	AGCTGGGGCG	CCCTCTGGTA	AGGTTGGGAA	3600
	GCCCTGCAAA	GTAAACTGGA	TGGCTTTCTT	GCCGCCAAGG	ATCTGATGGC	GCAGGGGATC	3660
10	AAGATCTGAT	CAAGAGACAG	GATGAGGATC	GTTTCGCATG	ATTGAACAAG	ATGGATTGCA	3720
	CGCAGGTTCT AATCGGCTGC	CCGGCCGCTT TCTGATGCCG	GGGTGGAGAG CCGTGTTCCG	GCTATTCGGC GCTGTCAGCG	TATGACTGGG CAGGGGCGCC	CACAACAGAC CGGTTCTTTT	3780 3840
15	TGTCAAGACC	GACCTGTCCG	GTGCCCTGAA	TGAACTGCAG	GACGAGGCAG	CGCGGCTATC	3900
	GTGGCTGGCC	ACGACGGGCG	TTCCTTGCGC	AGCTGTGCTC	GACGTTGTCA	CTGAAGCGGG	3960
20	AAGGGACTGG	CTGCTATTGG	GCGAAGTGCC	GGGGCAGGAT	CTCCTGTCAT	CTCACCTTGC	4020
20	TCCTGCCGAG	AAAGTATCCA	TCATGGCTGA	TGCAATGCGG	CGGCTGCATA	CGCTTGATCC	4080
	GGCTACCTGC	CCATTCGACC	ACCAAGCGAA	ACATCGCATC	GAGCGAGCAC	GTACTCGGAT	4140
25	GGAAGCCGGT	CTTGTCGATC	AGGATGATCT	GGACGAAGAG	CATCAGGGGC	TCGCGCCAGC	4200
	CGAACTGTTC	GCCAGGCTCA	AGGCGCGCAT	GCCCGACGGC	GAGGATCTCG	TCGTGACCCA	4260
20	TGGCGATGCC	TGCTTGCCGA	ATATCATGGT	GGAAAATGGC	CGCTTTTCTG	GATTCATCGA	4320
30	CTGTGGCCGG	CTGGGTGTGG	CGGACCGCTA	TCAGGACATA	GCGTTGGCTA	CCCGTGATAT	4380
	TGCTGAAGAG	CTTGGCGGCG	AATGGGCTGA	CCGCTTCCTC	GTGCTTTACG	GTATCGCCGC	4440
35	TCCCGATTCG	CAGCGCATCG	CCTTCTATCG	CCTTCTTGAC	GAGTTCTTCT	GAGCGGGACT	4500
	CTGGGGTTCG	AAATGACCGA	CCAAGCGACG	CCCAACCTCC	AGAAAAAGGG	GGGAATGAAA	4560
40	GACCCCACCT	GTAGGTTTGG	CAAGCTAGCT	TAAGTAACGC	CATTTTGCAA	GGCATGGAAA	4620
40	AATACATAAC	TGAGAATAGA	GAAGTTCAGA	TCAAGGTCAG	GAACAGATGG	AACAGCTGAA	4680
	TATGGGCCAA	ACAGGATATC	TGTGGTAAGC	AGTTCCTGCC	CCGGCTCAGG	GCCAAGAACA	4740
45	GATGGAACAG	CTGAATATGG	GCCAAACAGG	ATATCTGTGG	TAAGCAGTTC	CTGCCCCGGC	4800
	TCAGGGCCAA	GAACAGATGG	TCCCCAGATG	CGGTCCAGCC	CTCAGCAGTT	TCTAGAGAAC	4860
50	CATCAGATGT	TTCCAGGGTG	CCCCAAGGAC	CTGAAATGAC	CCTGTGCCTT	ATTTGAACTA	4920
50	ACCAATCAGT	TCGCTTCTCG	CTTCTGTTCG	CGCGCTTCTG	CTCCCCGAGC	TCAATAAAAG	4980
	AGCCCACAAC	CCCTCACTCG	GGGCGCCAGT	AATCTGCTGC	TTGCAAACAA	AAAAACCACC	5040
55	GCTACCAGCG	GTGGTTTGTT	TGCCGGATCA	AGAGCTACCA	ACTCTTTTTC	CGAAGGTAAC	5100
	TGGCTTCAGC	AGAGCGCAGA	тассааатас	TGTCCTTCTA	GTGTAGCCGT	AGTTAGGCCA	5160
<b>60</b>	CCACTTCAAG	AACTCTGTAG	CACCGCCTAC	ATACCTCGCT	CTGCTAATCC	TGTTACCAGT	5220
60	GGCTGCTGCC	AGTGGCGATA	AGTCGTGTCT	TACCGGGTTG	GACTCAAGAC	GATAGTTACC	5280
	GGATAAGGCG	CAGCGGTCGG	GCTGAACGG	GGGTTCGTGC	CACACAGCCCA	GCTTGGAGCG	5340
65	AACGACCTAC	ACCGAACTGA	GATACCTACA	A GCGTGAGCAT	TGAGAAAGCG	CCACGCTTCC	5400

	CGAAGGGAGA AAGGCGGACA GGTATCCGGT AAGCGGCAGG GTCGGAACAG GAGAGCGCAC	5460
	GAGGGAGCTT CCAGGGGGAA ACGCCTGGTA TCTTTATAGT CCTGTCGGGT TTCGCCACCT	5520
5	CTGACTTGAG CGTCGATTTT TGTGATGCTC GTCAGGGGGG CGGAGCCTAT GGAAAAACGC	5580
	CAGCAACGCC GAGA	5594
10	(2) INFORMATION FOR SEQ ID NO:30:	,
10	(i) SEQUENCE CHARACTERISTICS:	
	(A) LENGTH: 6561 base pairs (B) TYPE: nucleic acid	
15	(C) STRANDEDNESS: single (D) TOPOLOGY: linear	
	(ii) MOLECULE TYPE: DNA (genomic)	
20		
20	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:30:	
	GATCCCGGG TCGACCCGGG TCGACCCTGT GGAATGTGTG TCAGTTAGGG TGTGGAAAGT	60
25	CCCCAGGCTC CCCAGCAGGC AGAAGTATGC AAAGCATGCA TCTCAATTAG TCAGCAACCA	120
	GGTGTGGAAA GTCCCCAGGC TCCCCAGCAG GCAGAAGTAT GCAAAGCATG CATCTCAATT	180
30	AGTCAGCAAC CATAGTCCCG CCCCTAACTC CGCCCATCCC GCCCCTAACT CCGCCCAGTT	240
	CCGCCCATTC TCCGCCCCAT GGCTGACTAA TTTTTTTTAT TTATGCAGAG GCCGAGGCCG	300
35	CCTCGGCCTC TGAGCTATTC CAGAAGTAGT GAGGAGGCTT TTTTGGAGGC CTAGGCTTTT	360
	GCAAAAAGCT TCACGCTGCC GCAAGCACTC AGGGCGCAAG GGCTGCTAAA GGAAGCGGAA	420
	CACGTAGAAA GCCAGTCCGC AGAAACGGTG CTGACCCCGG ATGAATGTCA GCTACTGGGC	480
40	TATCTGGACA AGGGAAAACG CAAGCGCAAA GAGAAAGCAG GTAGCTTGCA GTGGGCTTAC	540
	ATGGCGATAG CTAGACTGGG CGGTTTTATG GACAGCAAGC GAACCGGAAT TGCCAGCTGG	600
45	GGCGCCCTCT GGTAAGGTTG GGAAGCCCTG CAAAGTAAAC TGGATGGCTT TCTTGCCGCC	660
43	AAGGATCTGA TGGCGCAGGG GATCAAGATC TGATCAAGAG ACAGGATGAG GATCGTTTCG	720
	CATGATTGAA CAAGATGGAT TGCACGCAGG TTCTCCGGCC GCTTGGGTGG AGAGGCTATT	780
50	CGGCTATGAC TGGGCACAAC AGACAATCGG CTGCTCTGAT GCCGCCGTGT TCCGGCTGTC	840
	AGCGCAGGGG CGCCCGGTTC TTTTTGTCAA GACCGACCTG TCCGGTGCCC TGAATGAACT	900
55	GCAGGACGAG GCAGCGCGC TATCGTGGCT GGCCACGACG GGCGTTCCTT GCGCAGCTGT	960
	GCTCGACGTT GTCACTGAAG CGGGAAGGGA CTGGCTGCTA TTGGGCGAAG TGCCGGGGCA	1020
	GGATCTCCTG TCATCTCACC TTGCTCCTGC CGAGAAAGTA TCCATCATGG CTGATGCAAT	1080
60	GCGGCGGCTG CATACGCTTG ATCCGGCTAC CTGCCCATTC GACCACCAAG CGAAACATCG	1140
	CATCGAGCGA GCACGTACTC GGATGGAAGC CGGTCTTGTC GATCAGGATG ATCTGGACGA	1200
65	AGAGCATCAG GGGCTCGCGC CAGCCGAACT GTTCGCCAGG CTCAAGGCGC GCATGCCCGA	1260
	CGGCGAGGAT CTCGTCGTGA CCCATGGCGA TGCCTGCTTG CCGAATATCA TGGTGGAAAA	1320

	•						
	TGGCCGCTTT	TCTGGATTCA	TCGACTGTGG	CCGGCTGGGT	GTGGCGGACC	GCTATCAGGA	1380
	CATAGCGTTG	GCTACCCGTG	ATATTGCTGA	AGAGCTTGGC	GGCGAATGGG	CTGACCGCTT	1440
5	CCTCGTGCTT	TACGGTATCG	CCGCTCCCGA	TTCGCAGCGC	ATCGCCTTCT	ATCGCCTTCT	1500
	TGACGAGTTC	TTCTGAGCGG	GACTCTGGGG	TTCGAAATGA	CCGACCAAGC	GACGCCCAAC	1560
10	CTGCCATCAC GTTTTCCGGG	GAGATTTCGA ACGCCGGCTG	TTCCACCGCC GATGATCCTC	GCCTTCTATG CAGCGCGGGG	AAAGGTTGGG ATCTCATGCT	CTTCGGAATC GGAGTTCTTC	1620 1680
	GCCCACCCCG	GAATTCGTAA	TCTGCTGCTT	GCAAACAAAA	AAACCACCGC	TACCAGCGGT	1740
15	GGTTTGTTTG	CCGGATCAAG	AGCTACCAAC	TCTTTTTCCG	AAGGTAACTG	GCTTCAGCAG	1800
	AGCGCAGATA	CCAAATACTG	TCCTTCTAGT	GTAGCCGTAG	TTAGGCCACC	ACTTCAAGAA	1860
20	CTCTGTAGCA	CCGCCTACAT	ACCTCGCTCT	GCTAATCCTG	TTACCAGTGG	CTGCTGCCAG	1920
20	TGGCGATAAG	TCGTGTCTTA	CCGGGTTGGA	CTCAAGACGA	TAGTTACCGG	ATAAGGCGCA	1980
	GCGGTCGGGC	TGAACGGGGG	GTTCGTGCAC	ACAGCCCAGC	TTGGAGCGAA	CGACCTACAC	2040
25	CGAACTGAGA	TACCTACAGC	GTGAGCATTG	AGAAAGCGCC	ACGCTTCCCG	AAGGGAGAAA	2100
	GGCGGACAGG	TATCCGGTAA	GCGGCAGGGT	CGGAACAGGA	GAGCGCACGA	GGGAGCTTCC	2160
20	AGGGGGAAAC	GCCTGGTATC	TTTATAGTCC	TGTCGGGTTT	CGCCACCTCT	GACTTGAGCG	2220
30	TCGATTTTTG	TGATGCTCGT	CAGGGGGGCG	GAGCCTATGG	AAAAACGCCA	GCAACGCCGA	2280
	GATGCGCCGC	CTCGAGTACA	CCTGCGTCAT	GCTGAGACCC	TCAAGCCTCA	CTAAAAGGGT	2340
35	CCCTGCCTAG	TTCTGTTTAC	TAATCTGCCT	TATTCTGTTT	TTGTTCCCAT	GTTAAAGATA	2400
	GAGTAAATGC	AGTATTCTCC	ACATAGAGAT	ATAGACTTCT	GAAATTCTAA	GATTAGAATT	2460
40	ATTTACAAGA	AGAAGTGGGG	AATGAAGAAT	AAAAAATTAC	TGGCCTCTTG	TGAGAACATG	2520
40	AACTTTCACC	TCGGAGCCCA	CCCCCTCCCA	TCTGGAAAAC	ATACTTGAGA	AAAACATTTT	2580
	CTGGAACAAC	CACAGAATGT	TTCAACAGGC	CAGATGTATT	GCCAAACACA	GGATATGACT	2640
45	CTTTGGTTGA	GTAAATTTGT	GGTTGTTAAA	CTTCCCCTAT	TCCCTCCCCA	TTCCCCCTCC	2700
	CAGTTTGTGG	TTTTTTCCTT	TAAAAGCTTG	TGAAAAATTT	GAGTCGTCGT	CGAGACTCCT	2760
50	CTACCCTGTG	CAAAGGTGTA	TGAGTTTCGA	CCCCAGAGCT	CTGTGTGCTT	TCTGTTGCTG	2820
50	CTTTATTTCG	ACCCCAGAGC	TCTGGTCTGT	GTGCTTTCAT	GTCGCTGCTT	TATTAAATCT	2880
	TACCTTCTAC	ATTTTATGTA	TGGTCTCAGT	GTCTTCTTGG	GTACGCGGCT	GTCCCGGGAC	2940
55	TTGAGTGTCT	GAGTGAGGGT	CTTCCCTCGA	GGGTCTTTCA	TTTGGTACAT	GGGCCGGGAA	3000
	TTCGAGAATC	: TTTCATTTGG	TGCATTGGCC	GGGAATTCGA	AAATCTTTCA	TTTGGTGCAT	3060
60	TGGCCGGGAA	ACAGCGCGAC	CACCCAGAGG	TCCTAGACCC	CACTTAGAGGT	AAGATTCTTT	3120
60	GTTCTGTTTT	GGTCTGATGT	CTGTGTTCTG	ATGTCTGTGT	TCTGTTTCTA	AGTCTGGTGC	3180
	GATCGCAGTI	TCAGTTTTGC	GGACGCTCAG	TGAGACCGCG	CTCCGAGAGG	GAGTGCGGGG	3240
65	TGGATAAGGA	TAGACGTGTC	CAGGTGTCCA	CCGTCCGTTC	GCCCTGGGAG	ACGTCCCAGG	3300

	AGGAACAGGG	GAGGATCAGG	GACGCCTGGT	GGACCCCTTT	GAAGGCCAAG	AGACCATTTG	3360
	GGGTTGCGAG	ATCGTGGGTT	CGAGTCCCAC	CTCGTGCCCA	GTTGCGAGAT	CGTGGGTTCG	3420
5	AGTCCCACCT	CGTGTTTTGT	TGCGAGATCG	TGGGTTCGAG	TCCCACCTCG	CGTCTGGTCA	3480
	CGGGATCGTG	GGTTCGAGTC	CCACCTCGTG	TTTTGTTGCG	AGATCGTGGG	TTCGAGTCCC	3540
10	ACCTCGCGTC TGGCCGGCCT			CGAGTCCCAC TCTGGTTTCT			3600 3660
	TCCGCTCTTG	TTGTGACTAC	TGTTTTTCTA	AAAATGGGAC	AATCTGTGTC	CACTCCCCTT	3720
15	TCTCTGACTC	TGGTTCTGTC	GCTTGGTAAT	TTTGTTTGTT	TACGTTTGTT	TTTGTGAGTC	3780
13	GTCTATGTTG	TCTGTTACTA	TCTTGTTTTT	GTTTGTGGTT	TACGGTTTCT	GTGTGTGTCT	3840
	TGTGTGTCTC	TTTGTGTTCA	GACTTGGACT	GATGACTGAC	GACTGTTTTT	AAGTTATGCC	3900
20	TTCTAAAATA	AGCCTAAAAA	TCCTGTCAGA	TCCCTATGCT	GACCACTTCC	TTTCAGATCA	3960
	ACAGCTGCCC	TGCCTCCCAC	TCCAACTCCA	GAGAGCAGCC	AGCGGGTCAC	AGTGGTCCCG	4020
25	CCCATGAACC	TGGAGCCTAG	GGAAAAATGA	GCTCGGAAAT	CCGGAGCAAA	TGAGGAGTGG	4080
23	TCCCTGAGAA	GTCAGTGGCC	TAAATGTTGT	GGCTGCTGAA	GCAAAAGAAG	AGGAGGCTGT	4140
	TCGAGTAGCC	GGCCAAGAGC	GCCGCGGGTT	CCCAGGCAGC	TTCTCATTCC	CCTGTCCCTC	4200
30	CCATCCCGTC	TCTTGTTAAC	AGAAAAACTG	CTTTCACTTT	GAGATATGAG	TGGCCCGATA	4260
	CAGCCAGCTG	TGAGAGCTGT	ACTCCCTTCC	CTGCCCCACG	TGTTTTCTCT	TCTCAGGCGA	4320
35	CCCCTCCCTG	AGCTGCTGGC	AGTGAGTCTG	TTCTAAGCTC	CAGTGAGGGA	GGCATCCGCC	4380
33	CACTTGGGGC	TTCTGTCCAA	GGTAAGGAGC	ACCTGTGAGT	CTAACTGCCA	GGCTCTGATG	4440
	GGGGTCTCGT	CTCTGTGGGA	CTAGAAAGTG	TCCCAACAAT	CTGACCAAGG	TAACAGGAAG	4500
40	TTAAGACAAA	GACAGAGACC	AAAGTCAGAA	TCAGAGCTGT	GCTGTGAGAC	AAAAAGATAA	4560
	AAAAAATAAA	ATGCTGGCCA	CAAAAGTCAG	GAAAACTAGA	AAACTTAGAT	AGTACCTGGC	4620
45	AACAAAAGAA	AGCTTTTGGC	TAAAGATCAA	CGTGTATACT	GTAAAGAAAA	TGAGCACTGG	4680
43	GTGAGAGACT	GCCCCAACAA	AAAGAAGAGG	AGCCCCCCTC	ATGACCAAAC	CCTTCACCTG	4740
	TTCGTGGCTA	AAAGTAAAGA	GATAACAAAA	GGGGTGCTAA	CACAGAAGCT	GAGTCCTTAA	4800
50	AAGAGTCCGG	TGGCCTACCT	GTTGAAGCAG	CTAAAAAAGA	GACTGTGTTT	CATACTCCTC	4860
	CACTGACCAG	TGCAAAACAA	GCTAAAAAGT	TCCTGGGCAC	TGCGGGCTTT	TGCAGATTGT	4920
55	GGATTCCAGG	TTTTGCTGAG	TTAAAGAGAT	AAACAGCCCT	TCGTATAGAA	AAATAAAAA	4980
55	CAACCTTGGA	TGTCCTTGGA	TGCTATTGAG	ACTGCCCTAA	TGTTGTCCCC	AGCTATGGGA	5040
	CTCCTAGATG	TGACTGAGAA	CAAAGGTATT	GCCAAAGAAG	TTCTTACTCA	GAGATTGGGA	5100
60	CCCTGAAAAA	GACCTGTGGC	ATACTTGTAA	GAAATTAGAC	CTGGTGGCTG	TAAGATGGCC	5160
	TGCTTGTCTG	CACATAGTGG	CTTCTGGTCA	AGGACGCAGA	TAAATTGACT	CTGAGACAAA	5220
65	ACTTGGCACA	TGTCCTAGAA	AGTGTGGTTC	AGCCCCCATG	ACCGATGGCT	GACTAACGCT	5280
UJ	CTTGAAAACA	TTATCCAACT	GTTCCCCTGA	CCGATGGACA	CATTGTCAGA	GCTTTTTTTG	5340

	ACTGAACGAG TGACCTTCGC TCCCCCTGCT ATCCTCGATC TCACTACTGC CTGAGACTTC	5400
5	ACCTACTCAT CATTGTGCTG ACATTCTGGC AGAAGAAACT CATACTCGAA ATGATCTGAA	5460
3	GGATCAGATC AGCCTTGGCC TGAGAGTTTG AGCTGGTACA CGGATGGCAG TAGCCTGGAG	5520
10	GTTAAGGGTA AGCGGAAGGC GGGGACAGCA GTGCAGTGGT GGACAGAAAG CAAGTGATCT AGGCCAGCAG CCTCCCTAAA GGGACTTCAG CCCACAAAGC CAAACTTGTG GCTTTAATAC	5580 5640
10	AAGCTCTGTA AATGGTAAAA AAAAAAAAGT CTACACGGAC AGCAGGTATG CTCTTGCCAC	5700
	TGTACAGAGC AATATACAGA CAAAGAGAAC TGTTGACATC TGCAGAGAAA GACCTAAGAT	5760
15	GCTGTGGCTA AAAGAAATCA GATGGCAAAT CTAACCGCCC AGGCATCCTA AAGAGCAATG	5820
	ATCCTGACAG TCTGAAGACT ATCAAGTTAT AGACAAATTA AGACTGGTAA AAAAAACCCT	5880
20	GTATAAAATA GTAAAAACTG AAAAAAGAAA ACTAGTCCTC TCATGAGAAG ACAGACCTGA	5940
20	CATCTACTGA AAAATAGACT TTACTGGAAA AAATATGTGT ATGAATACCT TCTAGTTTTT	6000
	GTGAACGTTC TCAAGATGGA TAAAAGCTTT TCCTTGTAAA ACGAGACTGA TCAGATAGTC	6060
25	ATCAAGAAGA TTGTTAAAGA AAATTTTCCA AGGTTCGGAG TGCCAAAAGC AATAGTGTCA	6120
	GATAATGGTC CTGCCTTTGT TGCCCAGGTA AGTCAGGGTG TGGCCAAGTA TTTAGAGGTC	6180
20	AAATGAAAAT TCCATTGTGT GTACAGACCT CAGAGCTCAG GAAAGATAAA AAAGAATAAA	6240
30	TAAAACTCTA AACAGACCTT GACAAAATTA ATCCTAGAGA CTGGCACAGA CTTACTTGGT	6300
	ACTCCTTCCC CTTGCCCTAT TTAGAACTGA GAATACTCCC TCTTGATTCG GTTTTACTCT	6360
35	TTTTAAGATC CTTTATGGGG CTCCTATGCC ATCACTGTCT TAAATGATGT GTTTAAACCT	6420
	ATGTTGTTAT AATAATGATC TATATGTTAA GTTAAAAGGC TTGCAGGTGG TGCAGAAAGA	6480
4.0	AGTCTGGTCA CAACTGGCTA CAGTGAACAA GCTGGGTACC CCAAGGACAT CTTACCAGTT	6540
40	CCAGCCAGAG ATCTGATCTA C	6561
	(2) INFORMATION FOR SEQ ID NO:31:	
45	(i) SEQUENCE CHARACTERISTICS:	
	(A) LENGTH: 55 base pairs (B) TYPE: nucleic acid	
50	<pre>(C) STRANDEDNESS: single (D) TOPOLOGY: linear</pre>	
50	(ii) MOLECULE TYPE: DNA (genomic)	
55	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:31:	
	GACTAACCTT GATTCCACTG GAGCCGTATT ACCGCCATGC ATTAGTTATT AATAG	55
60	(2) INFORMATION FOR SEQ ID NO:32:	
υυ	(i) SEQUENCE CHARACTERISTICS:	
	(1) SEQUENCE CHARACTERISTICS:  (A) LENGTH: 47 base pairs  (B) TYPE: nucleic acid	
65	(B) TYPE: Nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear	

(ii) MOLECULE TYPE: DNA (genomic) 5 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:32: 47 GACTAACCTT GATTCCACTG GAGTAATTGC GGCTAGCGGA TCTGACG 10 (2) INFORMATION FOR SEQ ID NO:33: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 66 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single 15 (D) TOPOLOGY: linear (ii) MOLECULE TYPE: DNA (genomic) 20 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:33: GACTAACCTT GATTCCACTG GAGACACTTG ACCTCTACCG CGCCAGTCCT CCGATTGACT 60 25 66 GAGTCG (2) INFORMATION FOR SEQ ID NO:34: 30 (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 48 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear 35 (ii) MOLECULE TYPE: DNA (genomic) 40 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:34: GACTAACCTT GATTCCACTG GAGGGATCCG CGCCCATGAT TATTATCG 48 45 (2) INFORMATION FOR SEQ ID NO:35: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 55 base pairs 50 (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: DNA (genomic) 55 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:35: 60 GACTAACCTT GATTCCAGCA ATGTCATGGC TACAGGCTCC CGGACGTCCC TGCTC 55 (2) INFORMATION FOR SEQ ID NO:36: 65 (i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 48 base pairs

(B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: DNA (genomic) 5 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:36: 10 GACTAACCTT GATTCCAGCA ATGTTAGGAC AAGGCTGGTG GGCACTGG 48 (2) INFORMATION FOR SEQ ID NO:37: (i) SEQUENCE CHARACTERISTICS: 15 (A) LENGTH: 49 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear 20 (ii) MOLECULE TYPE: DNA (genomic) 25 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:37: 49 GACTAACCTT GATTCCACTG GAGGGTCGAC CCTGTGGAAT GTGTGTCAG (2) INFORMATION FOR SEQ ID NO:38: 30 (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 48 base pairs (B) TYPE: nucleic acid 35 (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: DNA (genomic) 40 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:38: 48 45 GACTAACCTT GATTCCACTG GAGAATCTCG TGATGGCAGG TTGGGCGT (2) INFORMATION FOR SEQ ID NO:39: (i) SEQUENCE CHARACTERISTICS: 50 (A) LENGTH: 54 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear 55 (ii) MOLECULE TYPE: DNA (genomic) (xi) SEQUENCE DESCRIPTION: SEQ ID NO:39: 60 GACTAACCTT GATTCCACTG AAGAGATTTT ATTTAGTCTC CAGAAAAAGG GGGG 54 (2) INFORMATION FOR SEQ ID NO:40: 65 (i) SEQUENCE CHARACTERISTICS:

5	<ul><li>(A) LENGTH: 50 base pairs</li><li>(B) TYPE: nucleic acid</li><li>(C) STRANDEDNESS: single</li><li>(D) TOPOLOGY: linear</li></ul>	•
3	(ii) MOLECULE TYPE: DNA (genomic)	
10	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:40:	
	GACTAACCTT GATTCCACTG AAGCCCCCAA ATGAAAGACC CCCGCTGACG	50
15	(2) INFORMATION FOR SEQ ID NO:41:	
20	(i) SEQUENCE CHARACTERISTICS:  (A) LENGTH: 49 base pairs  (B) TYPE: nucleic acid  (C) STRANDEDNESS: single  (D) TOPOLOGY: linear	
	(ii) MOLECULE TYPE: DNA (genomic)	
25		
	(xi) SEOUENCE DESCRIPTION: SEQ ID NO:41:	
30	GACTAACCTT GATTCCACTG GAGCCGGGAC GGAATTCGTA ATCTGCTGC	49
30		.,
	(2) INFORMATION FOR SEQ ID NO: 42:	
35	<ul> <li>(i) SEQUENCE CHARACTERISTICS:</li> <li>(A) LENGTH: 47 base pairs</li> <li>(B) TYPE: nucleic acid</li> <li>(C) STRANDEDNESS: single</li> <li>(D) TOPOLOGY: linear</li> </ul>	
40	(ii) MOLECULE TYPE: DNA (genomic)	
45	( ) OPOURNOR DESCRIPTION, SEC ID NO.42.	
43	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:42:	47
	GACTAACCTT GATTCCACTG GAGTTCTCGA GGCGGCGCAT CTCGGCG	3,
50	(2) INFORMATION FOR SEQ ID NO:43:	
	<ul><li>(i) SEQUENCE CHARACTERISTICS:</li><li>(A) LENGTH: 20 base pairs</li><li>(B) TYPE: nucleic acid</li><li>(C) STRANDEDNESS: single</li></ul>	
55	(D) TOPOLOGY: linear	
	(ii) MOLECULE TYPE: DNA (genomic)	
60		
	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:43:	
	CGCTCTAGAA CTAGTGGATC	20
65	(2) INFORMATION FOR SEQ ID NO:44:	

5	<ul> <li>(i) SEQUENCE CHARACTERISTICS:</li> <li>(A) LENGTH: 21 base pairs</li> <li>(B) TYPE: nucleic acid</li> <li>(C) STRANDEDNESS: single</li> <li>(D) TOPOLOGY: linear</li> </ul>	
	(ii) MOLECULE TYPE: DNA (genomic)	
10		
	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:44:	
		0.1
15	GTAATACGAC TCACTATAGG G	21
	(2) INFORMATION FOR SEQ ID NO:45:	
20	<ul> <li>(i) SEQUENCE CHARACTERISTICS:</li> <li>(A) LENGTH: 43 base pairs</li> <li>(B) TYPE: nucleic acid</li> <li>(C) STRANDEDNESS: single</li> <li>(D) TOPOLOGY: linear</li> </ul>	
25	(ii) MOLECULE TYPE: DNA (genomic)	
30	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:45:	
	CGATCCACTG GAGCTCGGAG CCCACCCCT CCCATCTAGA GGT	43
35	(2) INFORMATION FOR SEQ ID NO:46:	
40	<ul> <li>(i) SEQUENCE CHARACTERISTICS:</li> <li>(A) LENGTH: 43 base pairs</li> <li>(B) TYPE: nucleic acid</li> <li>(C) STRANDEDNESS: single</li> <li>(D) TOPOLOGY: linear</li> </ul>	
	(ii) MOLECULE TYPE: DNA (genomic)	
45		
	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:46:	
	CGTCCTCCTG GAGAGCACAG GGTAGAGGAG TCTCGACGGT CAG	43
50	(2) INFORMATION FOR SEQ ID NO:47:	
55	(i) SEQUENCE CHARACTERISTICS:  (A) LENGTH: 43 base pairs  (B) TYPE: nucleic acid  (C) STRANDEDNESS: single  (D) TOPOLOGY: linear	
60	(ii) MOLECULE TYPE: DNA (genomic)	
	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:47:	
65	CGCAACCCTG GAGACCTCTA GATGGGAGGG GGTGGGCTCC GAG	43

(2) INFORMATION FOR SEQ ID NO:48: (i) SEQUENCE CHARACTERISTICS: 5 (A) LENGTH: 43 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear 10 (ii) MOLECULE TYPE: DNA (genomic) 15 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:48: 43 GCAGGACCTG GAGCTGACCG TCGAGACTCC TCTACCCTGT GCT (2) INFORMATION FOR SEQ ID NO:49: 20 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 20 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single 25 (D) TOPOLOGY: linear (ii) MOLECULE TYPE: DNA (genomic) 30 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:49: 20 CGCTCTAGAA CTAGTGGATC 35 (2) INFORMATION FOR SEQ ID NO:50: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 21 base pairs 40 (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: DNA (genomic) 45 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:50: 50 GTAATACGAC TCACTATAGG G 21 (2) INFORMATION FOR SEQ ID NO:51: 55 (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 19 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear 60 (ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:51:

	TACGTATCGA TGGATCCGA (2) INFORMATION FOR SEQ ID NO:52:	19
5	<ul> <li>(i) SEQUENCE CHARACTERISTICS:</li> <li>(A) LENGTH: 19 base pairs</li> <li>(B) TYPE: nucleic acid</li> <li>(C) STRANDEDNESS: single</li> <li>(D) TOPOLOGY: linear</li> </ul>	
10	(ii) MOLECULE TYPE: DNA (genomic)	
15	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:52:	
	GGATCCATCG ATACGTAAG	19
20	(2) INFORMATION FOR SEQ ID NO:53:	
25	<ul> <li>(i) SEQUENCE CHARACTERISTICS:</li> <li>(A) LENGTH: 38 base pairs</li> <li>(B) TYPE: nucleic acid</li> <li>(C) STRANDEDNESS: single</li> <li>(D) TOPOLOGY: linear</li> </ul>	
30	(ii) MOLECULE TYPE: DNA (genomic)	
	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:53:	
35	GGCCGCTAAC TAATAGCCCA TTCTCCAAGG TACGTAGC	38
	(2) INFORMATION FOR SEQ ID NO:54:	
40	<ul> <li>(i) SEQUENCE CHARACTERISTICS:</li> <li>(A) LENGTH: 38 base pairs</li> <li>(B) TYPE: nucleic acid</li> <li>(C) STRANDEDNESS: single</li> <li>(D) TOPOLOGY: linear</li> </ul>	
45	(ii) MOLECULE TYPE: DNA (genomic)	
50	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:54:	
	TACGTACCTT GGAGAATGGG CTATTAGTTA GCGGCCGC	38
55	(2) INFORMATION FOR SEQ ID NO:55:	
60	<ul> <li>(i) SEQUENCE CHARACTERISTICS:</li> <li>(A) LENGTH: 55 base pairs</li> <li>(B) TYPE: nucleic acid</li> <li>(C) STRANDEDNESS: single</li> <li>(D) TOPOLOGY: linear</li> </ul>	
	(ii) MOLECULE TYPE: DNA (genomic)	
65		

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:55: GACTAACCTT GATTCCACTG GAGTTTTCTC TATTCTTCAT TCCCCACTTC TTCTT (2) INFORMATION FOR SEQ ID NO:56: 5 (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 60 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single 10 (D) TOPOLOGY: linear (ii) MOLECULE TYPE: DNA (genomic) 15 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:56: GACTAACCTT GATTCCACTG GAGAATCTGG ACCAATTCTA TATAAGCCTG TGAAAAATTT 60 20 (2) INFORMATION FOR SEQ ID NO:57: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 46 base pairs 25 (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear 30 (ii) MOLECULE TYPE: DNA (genomic) 35 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:57: 46 GACTAACCTT GATTCCACTG GAGAAGAAGA AGTGGGGAAT GAAGAA (2) INFORMATION FOR SEQ ID NO:58: 40 (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 51 base pairs (B) TYPE: nucleic acid(C) STRANDEDNESS: single 45 (D) TOPOLOGY: linear (ii) MOLECULE TYPE: DNA (genomic) 50 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:58: 51 GACTAACCTT GATTCCACTG GAGATCTCTA GATGGGAGGG GGTCTGGGCT C 55 (2) INFORMATION FOR SEQ ID NO:59: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 47 base pairs (B) TYPE: nucleic acid
(C) STRANDEDNESS: single 60 (D) TOPOLOGY: linear (ii) MOLECULE TYPE: DNA (genomic) 65

	$\cdot$	_
	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:59:	
5	GACTAACCTT GATTCCACTG GAGCTCGGAG CCCACCCCT CCCATCT	47
,	(2) INFORMATION FOR SEQ ID NO:60:	
10	<ul> <li>(i) SEQUENCE CHARACTERISTICS:</li> <li>(A) LENGTH: 47 base pairs</li> <li>(B) TYPE: nucleic acid</li> <li>(C) STRANDEDNESS: single</li> <li>(D) TOPOLOGY: linear</li> </ul>	
15	(ii) MOLECULE TYPE: DNA (genomic)	
20	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:60:	
	GACTAACCTT GATTCCACTG GAGGGAGGCC CTTATCTCAA AAATGTT	47
	(2) INFORMATION FOR SEQ ID NO:61:	
25	<ul><li>(i) SEQUENCE CHARACTERISTICS:</li><li>(A) LENGTH: 51 base pairs</li><li>(B) TYPE: nucleic acid</li><li>(C) STRANDEDNESS: single</li></ul>	
30	(D) TOPOLOGY: linear  (ii) MOLECULE TYPE: DNA (genomic)	
35	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:61:	
	GACTAACCTT GATTCCACTG GAGTCTAAGA ACATTTTTGA GATAAGGGCC T	51
40	(2) INFORMATION FOR SEQ ID NO:62:	
45	<ul> <li>(i) SEQUENCE CHARACTERISTICS:</li> <li>(A) LENGTH: 44 base pairs</li> <li>(B) TYPE: nucleic acid</li> <li>(C) STRANDEDNESS: single</li> <li>(D) TOPOLOGY: linear</li> </ul>	
	(ii) MOLECULE TYPE: DNA (genomic)	
50		
	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:62:	
55	GACTAACCTT GATTCCACTG GAGTCACAGG CTTATATAGT GAAA	4.4
	(2) INFORMATION FOR SEQ ID NO:63:	
60	(i) SEQUENCE CHARACTERISTICS:  (A) LENGTH: 46 base pairs  (B) TYPE: nucleic acid  (C) STRANDEDNESS: single  (D) TOPOLOGY: linear	
65	(ji) MOLECULE TYPE: DNA (genomic)	

	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:63:	
5	GACTAACCTT GATTCCCTGG AGACTGCACT GCTGTCCCCG CCTTCG	46
	(2) INFORMATION FOR SEQ ID NO:64:	
10	<ul> <li>(i) SEQUENCE CHARACTERISTICS:</li> <li>(A) LENGTH: 53 base pairs</li> <li>(B) TYPE: nucleic acid</li> <li>(C) STRANDEDNESS: single</li> <li>(D) TOPOLOGY: linear</li> </ul>	
15	(ii) MOLECULE TYPE: DNA (genomic)	
20	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:64:	
	GAGTAACCTT GATTCCCTGG AGATTTCTCA GACCCGGGTC GACCCTGTGG AAT	53
25	(2) INFORMATION FOR SEQ ID NO:65:	
23	(i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 44 base pairs	
	(B) TYPE: nucleic acid (C) STRANDEDNESS: single	
30	(D) TOPOLOGY: linear	
	(ii) MOLECULE TYPE: DNA (genomic)	
35		
	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:65:	
	GACTAACCTT GATTCCCTGG AGCTCGAGGC GGCGCATCTC GGCG	44
40	(2) INFORMATION FOR SEQ ID NO:66:	
45	(i) SEQUENCE CHARACTERISTICS:  (A) LENGTH: 47 base pairs  (B) TYPE: nucleic acid  (C) STRANDEDNESS: single  (D) TOPOLOGY: linear	
50	(ii) MOLECULE TYPE: DNA (genomic)	
55	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:66:	
-	GACTAACCTT GATTCCCTGA AGACCTGCGT CATGCTGAGA CCCTCAA	47
	(2) INFORMATION FOR SEQ ID NO:67:	
60	<ul><li>(i) SEQUENCE CHARACTERISTICS:</li><li>(A) LENGTH: 50 base pairs</li><li>(B) TYPE: nucleic acid</li><li>(C) STRANDEDNESS: single</li></ul>	
65	(D) TOPOLOGY: linear (ii) MOLECULE TYPE: DNA (genomic)	

. 5	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:67:	
3	GACTAACCTT GATTCCCTGA AGCGGCCAAT GCACCAAATG AAAGATTTTC	50
	(2) INFORMATION FOR SEQ ID NO:68:	
10	<ul> <li>(i) SEQUENCE CHARACTERISTICS:</li> <li>(A) LENGTH: 50 base pairs</li> <li>(B) TYPE: nucleic acid</li> <li>(C) STRANDEDNESS: single</li> <li>(D) TOPOLOGY: linear</li> </ul>	
15	(ii) MOLECULE TYPE: DNA (genomic)	
20		
	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:68:	
	CGCATCTTTT AATTAACTGG AGARAATTTT TYACAGGCTT ATATAGKAAA	50

PCT/US98/03918

15

20

25

30

1. A method for assembling a gene or gene vector comprising the steps of:

- a) designing at least 6 primers to produce at least three fragments in at least three separate polymerase chain reactions wherein each primer comprises at least one predetermined restriction endonuclease recognition site that recognizes a restriction endonuclease that cleaves at a distance from the recognition site, a sequence complementary to a template sequence for amplification, and bases positioned at the restriction endonuclease cleavage site that are selected to be complementary to only one other overhanging created from enzymatic cleavage of the fragments;
  - b) combining the primers with template nucleic acid and performing a gene amplification reaction to produce multiple copies of an amplified template fragment incorporating the restriction endonuclease recognition site;
  - c) digesting the amplified template fragments with one or more restriction endonucleases that recognize the restriction endonuclease recognition site of the primers to create overhanging termini wherein each overhanging termini is complementary to only one other overhanging termini on another fragment; and
  - d) combining the amplified and digested template fragments in a ligation reaction to produce a directionally ordered gene, nucleic acid fragment or gene vector.
  - 2. The method of claim 1 wherein the restriction endonuclease is at least one class IIS restriction endonuclease.
  - 3. The method of claim 2 wherein the class IIS restriction endonuclease is selected from the group consisting of: AlwI, Alw26I, BbsI, BbvI, BbvII, BpmI, BsmAI, BsmI, BsmBI, BspMI, BsrI, BsrDI, Eco57I, EarI, FokI, GsuI, HgaI, HphI, MboII, MnII, PleI, SapI, SfaNI, TaqII, Tth111II.
  - 4. The method of claim 1 wherein class II restriction endonuclease recognition sites, linkers, or adapters are not used to create the gene or gene vector.

- 5. The method of claim 1 wherein the product of the ligation reaction is introduced into prokaryotic or eukaryotic cells.
- 5 6. The method of claim 1 wherein at least one target nucleic acid sequence is chosen from the group consisting of: transcriptional regulatory sequences; genetic vectors; introns and/or exons; viral encapsidation sequences; integration signals intended for introducing nucleic acid molecules into other nucleic acid molecules; retrotransposon(s); VL30 elements; or multiple allelic forms of a sequence.

10

- 7. The method of claim 1 wherein the method is used to generate combinatorial libraries of a target sequence.
- 8. The method of claim 7 wherein the target sequence is part or all of a gene.

15

- 9. The method of claim 8 wherein the gene encodes a protein.
- 10. The method of claim 8 wherein the primers amplify allelic variants of part or all of a gene.

- 11. The method of claim 1 wherein the product of the ligation reaction is passed between eukaryotic cells using a virus particle, by cell fusion, or by transfection.
- 12. The method of claim 1 wherein the product of the ligation reaction is not introduced into prokaryotic cells.
  - 13. The method of claim 1 further combining at least one screening or selection step to select the products of the ligation reaction.
- 30 14. The method of claim 1 wherein the product of the ligation reaction is mutated during passage in cells in order to generate genetic diversity.

- 15. The method of claim 14 wherein the product of the ligation reaction is mutated byhomologous recombination during passage in cells.
- 16. The method of claim 1, wherein the method is used to isolate and identify regulatorysequences from a cell.
  - 17. The method of claim 11, wherein cells containing the product of the ligation reaction are selected for enhanced biological activity.
- 10 18. The method of claim 17, wherein the cells containing the product of the ligation reaction are selected for tissue-specific, hormone-specific or developmental-specific gene expression.
- 19. The method of claim 1 wherein the product of the ligation reaction is a circularized15 gene vector.
  - 20. A nucleic acid primer having a 5' and a 3' end to amplify a nucleic acid fragment for the ligation of at least two fragments comprising:
- a restriction endonuclease recognition site that recognizes a restriction endonuclease,
  wherein the restriction endonuclease cleaves at a distance from the recognition site and
  creates overhanging termini;
  - a sequence complementary to a template sequence to be amplified to produce the nucleic acid fragment;
- at least two nucleic acid bases positioned at the restriction endonuclease cleavage site

  25 and that form an overhanging terminus after cleavage by the restriction endonuclease,
  wherein the at least two nucleic acid bases are selected to be complementary to only one other
  overhanging terminus on another fragment of the ligation; and

an affinity handle on the 5' end of the primer.

The primer of claim 20 further comprising an anchor to provide stability to the restriction enzyme at the restriction enzyme recognition site.

- 22. A method for isolating and identifying promoters comprising the steps of:
- a) obtaining a vector comprising at least a portion of a promoter region from a retrovirus transposon LTR and having two non-complementary overhanging termini;
- b) designing at least two PCR primers to amplify at least one region of a retro-transposon LTR from template nucleic acid to produce at least one nucleic acid fragment wherein each primer comprises at least one predetermined restriction endonuclease recognition site that recognizes a restriction endonuclease that cleaves at a distance from the recognition site, a sequence complementary to a template sequence from a retrovirus transposon, and bases positioned at the restriction endonuclease cleavage site that are selected to be complementary to only one other overhanging terminus of the vector wherein the restriction endonuclease cleavage of the fragments;
  - c) combining the primers with template nucleic acid and performing a gene amplification reaction to produce multiple copies of an amplified template fragment incorporating the restriction endonuclease recognition site;
- d) digesting the amplified template fragments with one or more restriction
   endonuclease that recognize the restriction endonuclease recognition site of the primer
   to create overhanging termini; and
  - e) combining the amplified and digested template fragment in a ligation reaction with the vector to produce a gene vector with an intact LTR sequence.
  - 23. The method of claim 22 wherein the template nucleic acid is DNA or RNA.
  - 24. The method of claim 22 further comprising the step of sequencing the insert to identify the promoter sequence.
  - 25. Promoter sequences of SEQ ID NOS:2-13 identified using the methods of claim 22.
  - 26. The vector of SEQ ID NO:1.

25

Fig 1A

A.

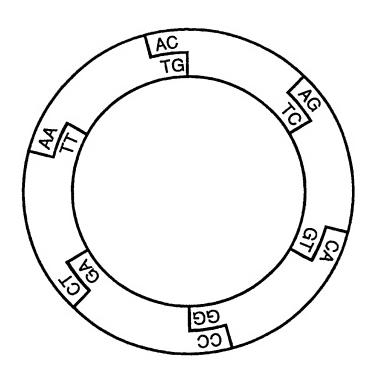
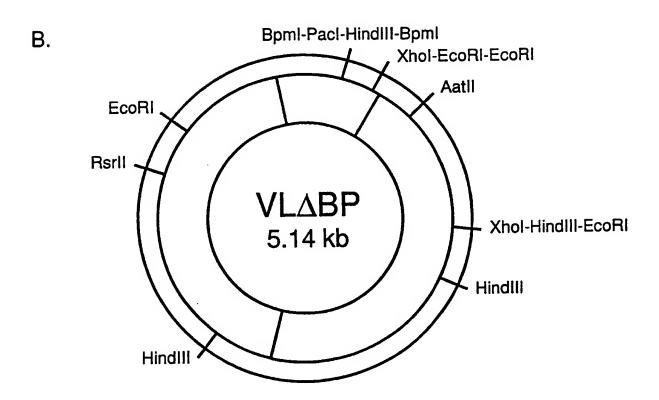
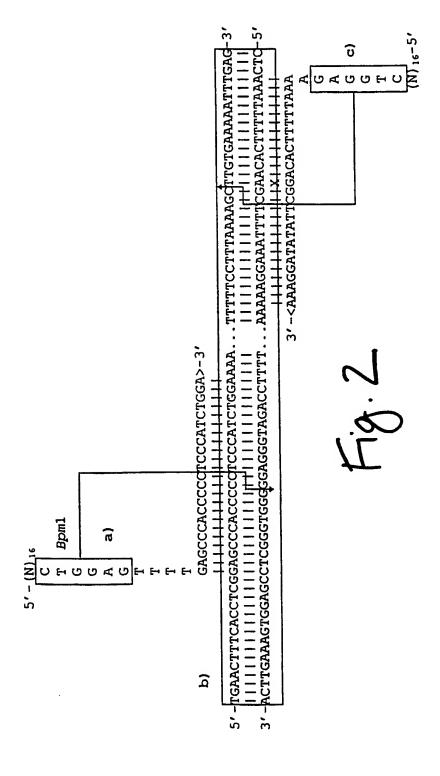


Fig 1B





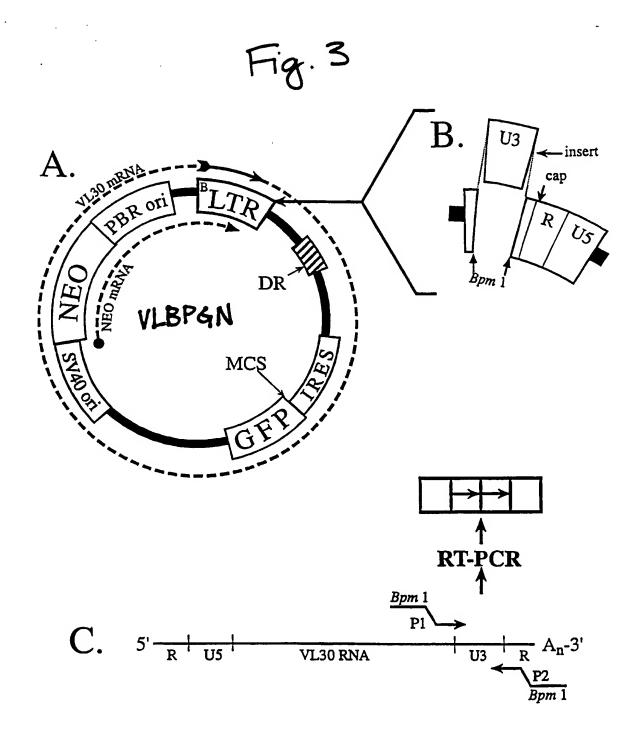


Fig. 4 A.

Genomic DNA or cellular RNA

Amplification of allelic parts via PCR or RT-PCR

Combine the parts in defined order using self-assembling genes

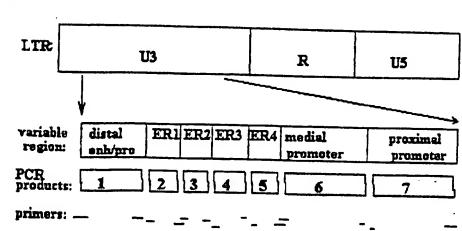
Grow constructs en masse

Transfect cells with constructs + replication competent retrovirus

Passage vectors that are expressed in mass cultures

Reisolate vectors after several passages

Fig. 4 B.



H.B. 5

									•
I.D.2	1.D.2 CCTCCCATCT AGAGGTTGTT CTCGGAACAC TCCTAAACTT TTCACCCCAA AACTCCTCAC CCTAAAGTTC GAAAAAACTG TTCCAAGAAC	AGAGGTTGTT	CTCGGAACAC	TCCTAAACTT	TTCACCCCAA	AACTCCTCAC	CCTAAAGTTC	GAAAAAACTG	TTCCAAGAAC
1.0.2	180 1.D.2 ATTITIONG TAAAGGCCTC CTAGAACAAC CTCAAAATGA CATTGCCAAA TGATAAGACA TGACTCCTTA GTTACGTAGG TTCCTTGATA T.D.2 ATTITIONG TAAAGGCCTC CTAGAACAAC CTCAAAAAAC ATTITIOAGA TAAAGGCTTC CTGGAACAAC CTCAAAATGA	TAAAGGCCTC	CTAGAACAAC	CTCAAAATGA	CATTGCCAAA TCTAGAAAAC	TGATAAGACA ATTTTGAGA	TGACTCCTTA	GTTACGTAGG CTGGAACAAC	180 TTCCTTGATA CTCAAAATGA
10.1	270 18.2 GARCATGACT CCTINGTIAE GRAGGITCCT TGATAGGACA TGACTCCTIA GITACGIAGA TICCTITGGI AGAACTCCCT AGTGAIGIAA 1.D.3 ACCAGGIACT CCTINGTIAE GRAGGITCCT TGATAGGACA TGACTCCTIA GITACAIAGA TICCTITGGE AGAACTCCCT AGTGAIGTAA 1.D.3 ACCAGGIACT CCTINGTIAE GRAGGITCCT TGATAGGACA TGACTCCTIA GITACAIAGA TICCTITGGE AGAACTCCCT AGTGAIGTAA	CCTTAGITAC	GTAGGTTCCT	TGATAGGACA TGATAGGACA	TGACTCCTTA TGACTCCTTA	gttacgtaga Gttacataga	meermeer meermee	AGAACTCCCT AGAACTCCCT	270 AGTGATGTAA AGTGATGTAA
1.0.2	121 271 1.D.2 ACTIGIACT TCCCTGCCCA GTTCTCCCC TTTGAGTTTT ACTATATANG C	TCCCTGCCCA	GFICTCCCCC	titicacititi Titicacititi	321 ACTATATAAG C ACTATATAAG C	400			

\	9	
Ĺ	7	)

36	TCGAACCTC TCGAACCTC TCGAACCTC TCGAACCCTC TCGAACCCTC TCGAACCCTC TCGAACCCTC TCGAACCTC TCGAACCTC TCGAACCTC TCGAACCCTC TCGAACCCTC TTGAACCCTC TTGAACCCTC	TCGAACCCTT	180 AATAATAGGA	AATAATAGGA	AATAATAGGA	AATAATAGGA	AATAATAGGA	AATAATAGGA	AATAATAGGA	AATGATAGGA
	AGAGATTGTT CCCAGAACAC TCCTGAACTC TTCACCCCAG AATGCATCC TGAACTCCTC ACCCTAGAGT TCGAACCCTC AGAGAGTGTT CCCAGAACAC TCCTGAACTC TTCACCCCAG AATGCATTCC TGAACTCCTC ACCCTAGAGT TCGAACCCTC AGAGATTGTT CCCAGAACAC TCCTGAACTC TTCACCCCAGA AATGCATTCC TGAACTCCTC ACCTAGAGT TCGAACCCTC AGAGATTGTT CCCAGAACAC TCCTGAACTC TTCACCCCAGA AATGCATTCC TGAACTCCTC ACCCTAGAGT TCGAACCCTC AGAGATTGT CCCAGAACAC TCCTGAACTC TTCACCCCAG AATGCATTCC TGAACTCCTC ACCTAGAGT TCGAACCCTC AGAGAGTGTT CCCAGAACAC TCCTGAACTC TTCACCCCAG AATGCATTCC TGAACCCCTC ACCTAGAGT TCGAACCCTC AGAGAGTGTT CCCAGAACAC TCCTGAACTC TTCACCCCAG AATGCATTCC TGAACTCCTC ACCTATAGT TCGAACCCTC AGAGAGTGTT CCCAGAACAC TCCTGAACTC TTCACCCCAG AATGCATTCC TGAACTCCTC ACCTATAGT TCAAACCCTC AGAGAGTGTT CCCAGAACAC TTCACCCCAG AATGCATTCC TGAACTCCTC ACCTATAGT TCAAAACCCTC AGAGAGTGTT CCCAGAACAC TTCACCCCAG AATGCATTCC TGAACTCCTC ACCTATAGT TCAAAACCCTC AGAGAGTGTT CCCAGAACAC TTCACCCCAG AATGCATTCC TGAACTCCTC ACCTATAGT TCAAAACCCTC AGAGAGTGTT CACCTAAACAC TCACCCCAGA AATGCATTCC TGAACTCCTC ACCTAAAGT TCAAACCCTC	TTCACCCCAG AATGCATTCC TGAACTCCTC ACCCTAGAGT TCGAACCCTT	180 GAGATAAGGG CCTCCTGGAA CAACCTCAGA ATGAACCGGG TACATTGCCA AATAATAGAA	TACATTECCA AATAATAGGA	CCTCCTGGAA CAACCTCAGA ATGAACCAGG TACATTGCCA AATAATAGGA	TACATTGCCA AATAATAGGA	TACATTGCCA AATAATAGGA	TACATTGCCA AATAATAGGA	CCTCCTGGAA CAACCTCAGA ATGAACCGGG TACATTGCCA AATAATAGGA	TACATTGCCA
	TGAACTCCTC	TGAACTCCTC	ATGAACCGGG	CCTCCTGGAA CAACCTCAGA ATGAACCGGA	ATGAACCAGG	CCTCCTGGAA CAACCTCAGA ATGAACCGGA	CCTCCTGGAA CAACCTCAGA ATGAACCAGG	CACATANGG CLICLISCAN CACCICAGA ATGAACCGGG	ATGAACCGGG	ATGAACCCCC
	TTCACCCCAG AATGCATGC TTCACCCCAG AATGCATTCC	AATGCATTCC	CAACCTCAGA	CAACCTCAGA	CAACCTCAGA	CAACCTCAGA	CAACCICAGA	CAACCTCAGA	CAACCTCAGA	CAACCTCAAA
	TTCACCCCAG TTCACCCCAG TTCACCCCAG TTCACCCCAG TTCACCCCAG TTCACCCCAG	TTCACCCCAG	CCTCCTGGAA	CCTCCTGGAA	CCICCIGGA	CCTCCTGGAA		CCICCIGGAA	CCTCCTGGAA	CCTCCTCGAA
	TCCTGAACTC	TCCTAAACTC	GAGATAAGGG		CACATAAGGG				GAGATAAGGG	GAGATAAGGG
	CCCAGAACAC CCCAGAACAC CCCAGAACAC CCCAGAACAC CCCAGAACAC CCCAGAACAC	CCCAGAACAC TCCTAAACTC	GAACAITITIT	GAACATTTTT	CAACATTITI	GAACATTTTT	GAACATTTTT	CARCATTTTT	GAACATCTTT	GAACATTTTT
	AGAGATTGTT AGAGAGTGTT AGAGAGTGTT AGAGAGTGTT AGAGAGTGTT AGAGAGTGTT AGAGAGTGTT AGAGAGTGTT AGAGAGTGTT	AGAGAGTGTT	ACTGTTCCAA GAACATTYTT	ACTGTTCCAA GAACATTTTT	ACTIGITACCAA	ACTGTTCCAA	CCAACTAAAG ACTGTTCCAA	CCAACTAAAG ACTGTTCCAA	CCARCINANG ACTUITICAN GRACATOTT	ACTUCAN
	CCTCCCATCT CCTCCCATCT CCTCCCATCT CCTCCCATCT CCTCCCATCT CCTCCCATCT CCTCCCATCT	CCTCCCATCT AGAGAGTGTT	91 CCAACTAAAG	CCAACTAAAG	CCAACTAAAG	CCAACTAAAG	CCAACTAAAG	CCAACTAAAG	CCARCINARIO	CCASTORAGE ACTION GAACATITIT GAGATAAGGG CCTCCTGGAA CAACCTCAAA ATGAACCGGG TACATITCCA AATGATAGAA
	47.00.8011	22		ın	vo r	~ ~	6	<u></u>	22	

Fig. 7

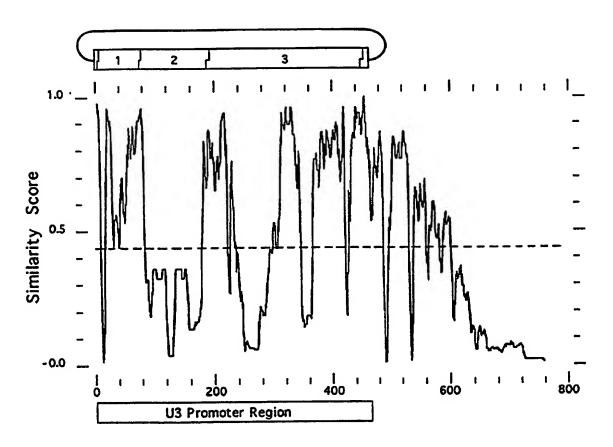
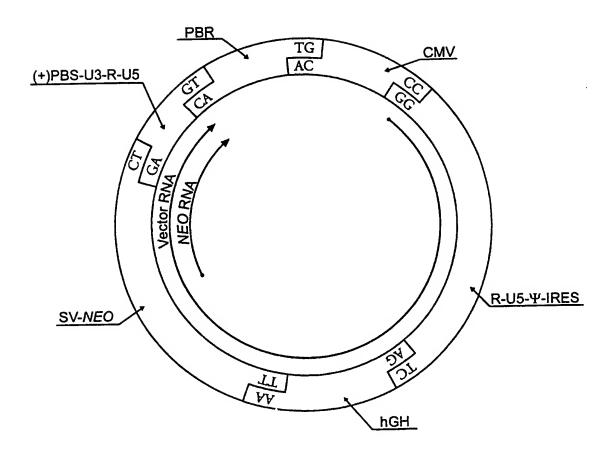


Fig. 8



## INTERNATIONAL SEARCH REPORT

i. iational Application No

A. CLASSIF IPC 6	FICATION OF SUBJECT MATTER C12N15/86 C12N15/10		-
According to	International Patent Classification(IPC) or to both national classific	ation and IPC	
B. FIELDS			
Minimum do IPC 6	cumentation searched (classification system followed by classificati C 12N	on symbols ₎	
Documentat	ion searched other than minimumdocumentation to the extent that s	such documents are included in the fields sea	rched
Electronic d	ata base consulted during the international search (name of data ba	se and, where practical, search terms used)	
C. DOCUME	ENTS CONSIDERED TO BE RELEVANT		
Category	Citation of document, with indication, where appropriate, of the re-	levant passages	Relevant to claim No.
Х	PADGETT K A ET AL: "Creating se junctions independent of restric in PCR cloning" GENE.		1,2, 4-14, 19-21
	vol. 168, no. 1, 2 February 1996 page 31-35 XP004042930	,	
Υ	see the whole document		3
Y	TOMIC, M. ET AL.: "A rapid and method for introducing specific into any position of DNA leaving positions unaltered" NUCLEIC ACIDS RESEARCH. vol. 18, no. 6, 1990, OXFORD GB page 1656 XP002069445 cited in the application see the whole document	mutations all other	3
X Furt	her documents are listed in the continuation of box C.	Patent family members are listed in	in annex.
"A" docum consid "E" earlier filing d "L" docum which catallo "O" docum other "P" docum later t	ent which may throw doubts on priority claim(s) or is cited to establish the publication date of another in or other special reason (as specified) lent referring to an oral disclosure, use, exhibition or means ent published prior to the international filing date but han the priority date claimed	"T" later document published after the inter- or priority date and not in conflict with cited to understand the principle or the invention  "X" document of particular relevance; the c- cannot be considered novel or cannot involve an inventive step when the do- "Y" document of particular relevance; the c- cannot be considered to involve an in- document is combined with one or m ments, such combination being obvio- in the art.  "3." document member of the same patent	the application but early underlying the stairmed invention to considered to coursent is taken alone claimed invention ventive step when the pre other such docurus to a person skilled family
	actual completion of theinternational search 26 June 1998	Date of mailing of the international sea	_{nur} , apar
Name and	mailing address of the ISA  European Patent Office. P.B. 5818 Patentlaan 2  NL - 2280 HV Rijswijk  Tel. (+31-70) 340-2040, Tx. 31 651 epo nl.	Authorized officer  Chambonnet, F	

## INTERNATIONAL SEARCH REPORT

In atlonal Application No
PCT/US 98/03918

artificial DNA splicing by directed ligation" NUCLEIC ACIDS RESEARCH, vol. 19, no. 24, 1991, OXFORD GB, pages 6757-6761, XP002069446 cited in the application see the whole document   CHAKRABORTY, A.K. ET AL.: "Synthetic retrotransposon vectors for gene therapy" FASEB JOURNAL., vol. 7, no. 10, July 1993, FOR EXPERIMENTAL BIOLOGY, BETHESDA, MD US, pages 971-977, XP002029486 see the whole document   P,X WO 97 28282 A (STRATAGENE INC) 7 August 1997	Relevant to claim No.
LEBEDENKO, E.N. ET AL.: "Method of artificial DNA splicing by directed ligation"  NUCLEIC ACIDS RESEARCH, vol. 19, no. 24, 1991, OXFORD GB, pages 6757-6761, XP002069446 cited in the application see the whole document  A CHAKRABORTY, A.K. ET AL.: "Synthetic retrotransposon vectors for gene therapy" FASEB JOURNAL., vol. 7, no. 10, July 1993, FOR EXPERIMENTAL BIOLOGY, BETHESDA, MD US, pages 971-977, XP002029486 see the whole document  P,X WO 97 28282 A (STRATAGENE INC) 7 August 1997  P,Y see the whole document  P,X HODGSON, C.P. ET AL.: "Self-assembling genes (SAGE): construction of complex vectors and combinatorial libraries without sub-cloning" CANCER GENE THERAPY, vol. 4, no. 6 conf. suppl. November 1997, page s27 XP002069448 see the whole document  P.X ZINK, A. M. ET AL.: "Transcriptional targeting with rescued LTRs: a hepatocyte promoter" CANCER GENE THERAPY, vol. 4, no. 6 conf. suppl., November 1997, page s28 XP002069449	1
artificial DNA splicing by directed ligation" NUCLEIC ACIDS RESEARCH, vol. 19, no. 24, 1991, OXFORD GB, pages 6757-6761, XP002069446 cited in the application see the whole document  A CHAKRABORTY, A.K. ET AL.: "Synthetic retrotransposon vectors for gene therapy" FASEB JOURNAL., vol. 7, no. 10, July 1993, FOR EXPERIMENTAL BIOLOGY, BETHESDA, MD US, pages 971-977, XP002029486 see the whole document  P,X WO 97 28282 A (STRATAGENE INC) 7 August 1997  P.Y see the whole document  P,X HODGSON, C.P. ET AL.: "Self-assembling genes (SAGE): construction of complex vectors and combinatorial libraries without sub-cloning" CANCER GENE THERAPY, vol. 4, no. 6 conf. suppl., November 1997, page s27 XP002069448 see the whole document  P.X ZINK, A. M. ET AL.: "Transcriptional targeting with rescued LTRs: a hepatocyte promoter" CANCER GENE THERAPY, vol. 4, no. 6 conf. suppl., November 1997, page s28 XP002069449	
retrotransposon vectors for gene therapy" FASEB JOURNAL., vol. 7. no. 10, July 1993, FOR EXPERIMENTAL BIOLOGY, BETHESDA, MD US, pages 971-977, XP002029486 see the whole document P,X WO 97 28282 A (STRATAGENE INC) 7 August 1997  P.Y see the whole document P,X HODGSON, C.P. ET AL.: "Self-assembling genes (SAGE): construction of complex vectors and combinatorial libraries without sub-cloning" CANCER GENE THERAPY, vol. 4. no. 6 conf. suppl. November 1997. page s27 XP002069448 see the whole document P.X ZINK, A. M. ET AL.: "Transcriptional targeting with rescued LTRs: a hepatocyte promoter" CANCER GENE THERAPY, vol. 4, no. 6 conf. suppl., November 1997, page s28 XP002069449	1
P,Y see the whole document  HODGSON, C.P. ET AL.: "Self-assembling genes (SAGE): construction of complex vectors and combinatorial libraries without sub-cloning" CANCER GENE THERAPY, vol. 4. no. 6 conf. suppl November 1997. page s27 XP002069448 see the whole document  P.X ZINK, A. M. ET AL.: "Transcriptional targeting with rescued LTRs: a hepatocyte promoter" CANCER GENE THERAPY, vol. 4, no. 6 conf. suppl., November 1997, page s28 XP002069449	
P,X  HODGSON, C.P. ET AL.: "Self-assembling genes (SAGE): construction of complex vectors and combinatorial libraries without sub-cloning"  CANCER GENE THERAPY, vol. 4. no. 6 conf. suppl November 1997. page s27 XP002069448 see the whole document   P.X  ZINK, A. M. ET AL.: "Transcriptional targeting with rescued LTRs: a hepatocyte promoter"  CANCER GENE THERAPY, vol. 4, no. 6 conf. suppl., November 1997, page s28 XP002069449	1.2, 4-14, 19-21
genes (SAGE): construction of complex vectors and combinatorial libraries without sub-cloning" CANCER GENE THERAPY, vol. 4. no. 6 conf. suppl November 1997. page s27 XP002069448 see the whole document  P.X  ZINK, A. M. ET AL.: "Transcriptional targeting with rescued LTRs: a hepatocyte promoter" CANCER GENE THERAPY, vol. 4, no. 6 conf. suppl., November 1997, page s28 XP002069449	3.
targeting with rescued LTRs : a hepatocyte promoter" CANCER GENE THERAPY, vol. 4, no. 6 conf. suppl., November 1997, page s28 XP002069449	1
	22

## INTERNATIONAL SEARCH REPORT

Information on patent family members

II iational Application No

 					FC1/U3	30/02310	
 Pa	tent document in search report		Publication date	Patent family member(s)		Publication date	
 WO	9728282	Α	07-08-1997	NONE			